

**AN AUTOMATIC METHOD FOR CLASSIFYING MEDICAL RESEARCHERS
INTO DOMAIN SPECIFIC SUBGROUPS**

by

Alfred A. Cecchetti

BS, Indiana University of Pennsylvania, 1972

MS, University of Pittsburgh, 1992

MSIS, University of Pittsburgh, 1996

Submitted to the Graduate Faculty of

The School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH
SCHOOL OF INFORMATION SCIENCES

This dissertation was presented
by

Alfred A. Cecchetti

It was defended on

April 20, 2009

and approved by

Robert A. Branch, MD, Professor of Medicine & Pharmacology, School of Medicine

Toni Carbo, PhD, Professor, Program in Library and Information Science

Sherry Koshman, PhD, Assistant Professor, Program in Library and Information Science

Dissertation Advisor: Ellen G. Detlefsen, DLS, Associate Professor, Program in Library
and Information Science

Copyright © by Alfred A. Cecchetti

2009

AN AUTOMATIC METHOD FOR CLASSIFYING MEDICAL RESEARCHERS INTO DOMAIN SPECIFIC SUBGROUPS

Alfred A. Cecchetti, MS, MSIS

University of Pittsburgh, 2009

Objective:

This dissertation developed an automatic classification procedure, as an example of a novel tool for an informationist, which extracts information from published abstracts, classifies abstracts into their “fields of study,” and then determines the researcher’s “field of study” and “level of activity.”

Method:

This dissertation compared a domain expert’s method of classification and an automatic classification procedure on a random sample of 101 medical researchers (derived from a potential list of 305 medical researchers) and their associated abstracts.

Design:

The study design is a retrospective, cross-sectional, inter-rater agreement study, designed to compare two classification methods (i.e., automatic classification procedure and domain expert). The study population consists of University of Pittsburgh, School of Medicine, Department of Medicine (DOM) professionals who (1) have published at

least one article listed in PubMed® as first or last author and/or (2) are the primary investigator for at least one grant listed in CRISP.

Main outcome measures:

Three outcome measures were derived from the domain expert's versus automatic categorization procedure: (1) an abstract's "field of study," (2) a researcher's "field of study" and (3) a researcher's "level of activity and field of study."

Results:

Kappa showed moderate agreement between automatic and domain expert classification for the abstracts' "field of study" (Kappa = 0.535, $n = 504$, $p < .000$).

Kappa showed moderate agreement between automatic and domain expert classification of the researcher's "field of study" (Kappa = 0.535, $n = 101$, $p < .000$).

Kappa showed good agreement between automatic and domain expert classification of the researcher's "level of activity and field of study" (Kappa = 0.634, $n = 101$, $p < .000$).

Conclusion:

The study suggests that an automatic library classification procedure can provide rapid classification of medical research abstracts into their "fields of study." The classification procedure can also process multiple abstracts' "fields of study" and classify their associated medical researchers into their "field of study" and "level of activity and field of study." The classification procedure, used as a tool by an informationist, can be used as the basis for new services.

TABLE OF CONTENTS

LIST OF TABLES.....	xii
LIST OF FIGURES.....	xiii
PREFACE.....	xv
1.0 INTRODUCTION	1
1.1 OVERVIEW	1
1.2 WHY IS CATEGORIZATION OF RESEARCHERS IMPORTANT?.....	1
1.3 WHO IN THE ACADEMIC MEDICAL LIBRARY WOULD USE THE CLASSIFICATION PROCEDURE?.....	3
1.4 WHAT FOUNDATIONS ARE THE PROCEDURE BUILT UPON?	3
1.5 HOW IS THE CLASSIFICATION SYSTEM TESTED?.....	4
2.0 PROBLEM STATEMENT	5
3.0 RESEARCH QUESTIONS.....	8
3.2 THE NATIONAL INSTITUTES OF HEALTH, TRANSLATIONAL RESEARCH, AND THE UNIVERSITY OF PITTSBURGH.....	9
3.3 BIG SCIENCE	9
3.4 THE CLASSIFICATION PROCEDURE	10

3.5 TEAM MEMBERS: THE INFORMATIONIST AND THE DOMAIN EXPERT	13
4.0 THE CLASSIFICATION PROCEDURE AS A RESOURCE	16
4.1 HOW CAN AUTOMATIC CLASSIFICATION SOLVE COMPLEX PROBLEMS	16
4.1.1 Solving the National Institutes of Health mentoring problem: An example of the classification procedure creating a novel new service.	17
4.1.2 Laying the foundation for library based push technology: An example of the classification system as an information filter.	22
4.2 WHY USE A DOCUMENT-BASED CLASSIFICATION PROCEDURE?	25
5.0 LITERATURE REVIEW SUMMARY:	29
5.1 PERSONALIZATION SYSTEMS AS A FILTER FOR GROUPS	29
5.2 AN EXAMINATION OF A DOCUMENT CLASSIFICATION SYSTEM: CITATION ANALYSIS	31
5.3 AN EXAMINATION OF THE DIFFERENT INFORMATION SEEKING BEHAVIOR OF ACADEMIC MEDICAL RESEARCHER SUBGROUPS	38
5.4 SUMMARY OF THE LITERATURE	46
6.0 CONCEPTS	48

6.1	FIRST AND LAST AUTHORS	49
6.2	PUBMED®	50
6.3	CRISP	50
6.4	ABSTRACT-TITLE	51
6.5	HIGHLY PRODUCTIVE MEMBERS	52
6.6	WORD/PHRASES IDENTIFIERS.....	52
6.7	SUBJECT-PREDICATE-OBJECT EXPRESSIONS	54
6.8	GROUPING MECHANISM.....	54
6.9	FIELD OF STUDY	56
6.10	LEVEL OF ACTIVITY AND FIELD OF STUDY	57
7.0	METHODOLOGY	58
7.1	STUDY DESIGN	59
7.2	FLOWCHART.....	60
7.3	INCLUSION CRITERIA	62
7.4	LIMITATIONS.....	62
7.5	ASSUMPTIONS	64
7.6	PUBMED® ABSTRACTS.....	65

7.7	CRISP ABSTRACTS.....	65
7.8	ABSTRACT-TITLE VARIABLE.....	66
7.9	DOCUMENT-AUTHOR CLASSIFICATION PROCEDURE FRAMEWORK	66
7.10	OUTCOME MEASURES.....	82
7.10.1	Three outcome measures or endpoints.....	82
7.10.2	Level of Activity	85
7.10.3	Current Time-Period.....	86
7.11	OPERATIONAL DEFINITION FOR SUMMATION RULES	87
7.12	SAMPLE SIZE CALCULATIONS	92
7.13	STATISTICS.....	93
7.14	STATISTICAL ASSUMPTIONS.....	95
8.0	ANALYSIS OF RESEARCH QUESTIONS	96
8.1	PILOT STUDY.....	97
8.2	FULL STUDY	98
8.3	INITIAL POPULATION	99
8.4	ACTIVE AND VERY ACTIVE CLASSIFICATION.....	102
8.5	RESULTS OF THE RESEARCH QUESTIONS.....	103

8.5.1 Can an automated procedure classify abstracts from academic medical researchers' publications and grants into a "field of study?" What is the level of agreement between the automated procedure and the results derived from a domain expert?	103
8.5.2 Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their "field of study?" What is the level of agreement between the automated procedure and the results derived from a domain expert?	105
8.5.3 Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their "level of activity and field of study?" What is the level of agreement between the automated procedure and the results derived from a domain expert? ..	108
8.6 PILOT STUDY VERSUS FINAL STUDY	112
8.7 STUDY CONCLUSION	113
9.0 FUTURE DIRECTION AND CONCLUSION	117
9.1 CLASSIFICATION AS A METHOD OF DETERMINING A MEDICAL RESEARCHER'S LIBRARY NEEDS	117
9.2 PREDICTING ACCEPTANCE OF A NEW SERVICE	122
9.3 DYNAMIC CLASSIFICATION	126
9.4 FURTHER INVESTIGATION.....	126

9.5	FINAL CONCLUSION	127
10.0	ADDENDUM A. IRB APPROVAL	130
11.0	ADDENDUM B. CLASSIFICATION CODE	131
11.1	FUNCTION FLOW LIST	131
11.2	SELECTED FUNCTIONS	133
12.0	BIBLIOGRAPHY	137

LIST OF TABLES

Table 7-1 Strength of agreement.....	94
Table 8-1 Count of active versus very active researcher.....	102
Table 8-2 Classification procedure versus random pairing – Question 1.....	103
Table 8-3 Kappa for domain expert versus random pairing – Question 1.....	104
Table 8-4 Classification procedure versus domain expert – Question 1.....	104
Table 8-5 Kappa value - Question 1	105
Table 8-6 Classification procedure versus random pairing - Question 2	106
Table 8-7 Domain expert versus random pairing - Question 2	106
Table 8-8 Classification procedure versus domain expert - Question 2.....	107
Table 8-9 Kappa value - Question 2	107
Table 8-10 Classification procedure versus random pairing - Question 3	110
Table 8-11 Domain expert versus random pairing - Question 3	110
Table 8-12 Classification procedure versus domain expert - Question 3.....	111
Table 8-13 Kappa value - Question 3	112
Table 8-14 Advantages and Disadvantages of the Classification Procedure ...	115
Table 9-1 Domain Landscape.....	118

LIST OF FIGURES

Figure 3-1 Subject-predicate-object expressions.....	11
Figure 3-2 Metadata consists of words or phrase that are “equivalent to”	12
Figure 7-1 Classification flowchart.....	60
Figure 7-2 The informationist discusses the characteristics of each "field of study"	67
Figure 7-3 The query string	68
Figure 7-4 Raw PubMed® Abstract and Title concatenated into one Field	69
Figure 7-5 The abstract + title field is processed	71
Figure 7-6 Abstract + Title processed into a table of words.....	72
Figure 7-7 Basic Researcher group compared to Non-Basic Researcher Group	73
Figure 7-8 Unique words in Target Group	74
Figure 7-9 The word “mice” in the abstract-title variable is equivalent to a basic researcher	75
Figure 7-10 Interesting words from the clinical outcome "field of study"	76
Figure 7-11 Frequency count of interesting words.....	77
Figure 7-12 The word “discharge” is a "related to" triplet.....	78
Figure 7-13 The word "discharge" occurs four times in abstract number 454.....	79
Figure 7-14 The triplets analyze each abstract.....	80
Figure 7-15 Abstract analysis chart	91
Figure 7-16 Medical researcher field of study chart.....	91

Figure 8-1 Medical Researchers who have published at least one PubMed® Abstract	99
Figure 8-2 Medical Researchers who have at least one CRISP abstract	100
Figure 8-3 Researcher published PubMed® abstracts and/or CRISP abstract	101
Figure 9-1 Future direction: Predictive instrument	122

PREFACE

I would like to thank Robert Branch without whom this dissertation would not be possible. Bob's insight, skills, and knowledge are incredible and he has my deep appreciation for making every workday a wonderful learning experience. I would like to thank the other members of my committee, Toni Carbo and Sherry Koshman, who provided brilliant advice and direction. I would also especially like to thank Ellen Detlefsen, my dear advisor, who always made time for my endless fountain of questions and concerns. Without Ellen's help, my doctorate would never have been completed. My very deep appreciation to a dear friend, Evelyn Perloff, who provided words of encouragement, infinite amounts of help and wisdom, and helped me refocus when I wondered if my doctorate would ever occur.

My love and thanks to my family: my brothers, Raymond Cecchetti, Michael Cecchetti, Jeffrey Cecchetti, and Daniel Cecchetti who were with me as I went from my cellar laboratory with frogs and snakes to the completion of my doctoral studies. My remarkable parents, my father Alfred E. Cecchetti and my mother Ida Cecchetti, who always encouraged and believed in me. In my parent's house, everyday is a holiday with endless amounts of love. Without their loving foundation, I never would have accomplished any of my goals.

My love and thanks to my children Aaron Cecchetti, Nicholas Cecchetti, and Matthew Cecchetti. My dearest sons have put up with my studies for their entire lives and finally I can close this path. My doctorate is my way of saying to my sons; I love you, never stop following your dreams, you have it within you to become anything you want. My love and thanks to my wife, Kim Cecchetti, for her support throughout my studies and making a wonderful home for our children. My wife spent the majority of her time raising our children as I studied, went to class, and became involved in endless projects. Without her maintaining our universe, I would have been lost.

1.0 INTRODUCTION

1.1 OVERVIEW

This dissertation explores the role of an academic medical library informationist. The extended role involves the informationist as a developer of specialized information procedures within a medical research setting. These procedures are derived by fusing together both library and informatics techniques.

Presented in this dissertation is a procedure that categorizes medical researchers by analyzing their published abstracts. An evaluation of the inter-rater agreement between the new procedure and a domain expert's manual procedure is discussed.

1.2 WHY IS CATEGORIZATION OF RESEARCHERS IMPORTANT?

The academic medical institution (defined for this dissertation as the University of Pittsburgh, School of Medicine, Department of Medicine, <http://www.dept-med.pitt.edu/index.aspx>; <http://www.dept-med.pitt.edu/divisions.html>) is involved in patient care and performs both basic and clinical research (Levine, 2008). Government

grants, which are very competitive, are a significant source of funding for academic medical institutions (Levine, 2008). In 2005, the National Institutes of Health (NIH) proposed a very important grant. This large and far-reaching grant program, called the Clinical and Translational Science Award (CTSA) program, has a focus on “Big Science” projects.

The University of Pittsburgh Department of Medicine, a CTSA grant recipient, is encouraged to assist medical researchers, especially translational (i.e., a blend of basic and clinical science) researchers, to develop “Big Science” projects. The academic medical library is charged with assisting the academic medical institution and the medical researchers.

A library procedure that classifies medical researchers into field of study and levels of activity would assist the institution in their search for translational researchers. Additionally, a library procedure that classifies medical researchers into field of studies could allow the library to tailor library services to that specific field of study (i.e., a specific “Big Science” project would receive specific library services). In the business world, financial institutions will classify individuals into levels of credit using a FICO® score (e.g., good, fair, poor) so as to provide specific services to specific groups. The classification procedure will provide the same type of function, identifying subgroups so as to provide specific services to that subgroup.

Understanding the common needs of each field of study may provide insight into how the academic medical library can provide innovative new services to the medical

research community and provide the library a “place at the table” in the Clinical and Translational Science Awards program as well as other large grant programs.

1.3 WHO IN THE ACADEMIC MEDICAL LIBRARY WOULD USE THE CLASSIFICATION PROCEDURE?

In 2000, Davidoff and Florance introduced the concept of the Informationist, a profession based in library science, charged with the crucial role of synthesizing, retrieving, and presenting information to those within the clinical disciplines. Oliver et al. (2008) and others have described the informationists as cross trained specialists with specific content knowledge and the ability to provide in-depth information services. Their belief was that the informationist would be uniquely qualified to apply their expertise to information problem solving in a specific domain (Oliver et al., 2008, p. 51).

Informationists participate within the academic medical institution as members of research teams (Oliver, 2005, p. 67) and, with their cross-training backgrounds, can provide new services to these teams.

1.4 WHAT FOUNDATIONS ARE THE PROCEDURE BUILT UPON?

A literature review is presented that examines other text classification procedures. In addition, concepts underlying the procedure are introduced and referenced. The domain

expert, using an informal key informant interview process (Quandt & Arcury, 1997, p. 277), provided the rules for the classification procedure.

1.5 HOW IS THE CLASSIFICATION SYSTEM TESTED?

The new classification procedure was compared to the manual classification procedure of a domain expert (i.e., a senior research investigator whose institutional position has been to guide, assist, and allocate resources to a wide range of senior and junior investigators within a specific division of that institution) and a determination of inter-rater agreement was presented. A pilot study was presented and analyzed to determine if the classification procedure was able to perform adequately. Based on the positive response from the pilot study, an evaluation of the final study is presented in this dissertation. In addition, the future direction of the classification procedure and the role of informationist as tool developer are proposed.

2.0 PROBLEM STATEMENT

In 2005, Dr. Elias Zerhouni, Director of the National Institutes of Health (NIH), stressed the need for the medical research community to “translate the remarkable scientific innovations we are witnessing into health gains for the nation.” (2005a, p.1621) and to this end, the NIH funded a new program, the Clinical and Translational Science Awards (CTSAs). These awards were designed to “advance the assembly of institutional academic ‘homes’ that can provide integrated intellectual and physical resources for the conduct of original clinical and translational science “(p. 1622).

This call to action to provide integrated resources to the clinical, basic, and translational researchers within the medical research community is being addressed by a number of non-physician disciplines, such as informatics (Berner, 2008), dentistry (Bertolami, 2008), and pharmacy (Figg, 2008). Similarly, the academic medical library’s mission (Medical Library Association, 2000) is to provide assistance and information resources.

The medical research community consists of subgroups of basic, clinical and translational researchers (Zerhouni, 2007). All members of the medical researcher subgroup have an interesting characteristic; they must publish information (Angell,

1986) about their interests, grants or ideas. As their careers progress, their publication “tail” gets longer and additional information about their interests are published. The greater the number of submitted works or grants the medical researcher publishes, the more active the researcher is considered in their subgroup.

To understand the integration of intellectual and physical resources for the research community, this dissertation suggests that one must understand the needs of each part of the community (i.e., the research subgroups, classified into their field of study). Understanding the activity level and field of study of a medical researcher will provide insight into novel new library services, services that might not be traditional for a medical academic library, but could be used in a new ways to support and enhance research.

An automatic method of classification that categorizes academic medical professional into fields of study would provide valuable insight into how the academic medical library can assist the roadmap of the National Institutes of Health, by providing insight into the specific needs of each field of study by integrating the needs of the disciplines involved. Additionally, understanding the common needs of each field of study may provide insight into how the academic medical library can provide innovative new services to the medical research community and provide the library a “place at the table” in the CTSA program.

This dissertation asks the question, “Can medical research published abstracts (categorized into their own field of study) categorize medical research professionals into their field of study?”

There may be differences in the categorization of very active researchers (i.e., those who publish many research articles each year) and active researchers (i.e., those who publish very few articles each year). This dissertation also asks the question, “Can the activity level for each researcher’s field of study be found by analyzing the number of published abstracts produced by that researcher?”

3.0 RESEARCH QUESTIONS

1. Can an automated procedure classify abstracts from academic medical researchers' publications and grants into a "field of study" (i.e., Basic, Clinical Outcomes, Clinical Trial, and Translational)? What is the level of agreement between the automated procedure and the results derived from a domain expert?
2. Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their "field of study" (i.e., Basic, Clinical Outcomes, Clinical Trial, and Translational)? What is the level of agreement between the automated procedure and the results derived from a domain expert?
3. Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their "level of activity and field of study" (i.e., Active and Very Active Basic, Active and Very Active Clinical Outcomes, Active and Very Active Clinical Trial, Active and Very Active Translational)? What is the level of agreement between the automated procedure and the results derived from a domain expert?

3.2 THE NATIONAL INSTITUTES OF HEALTH, TRANSLATIONAL RESEARCH, AND THE UNIVERSITY OF PITTSBURGH

This dissertation explores the use of an automatic library based classification system to categorize a medical researcher population located at a large medical university, specifically the University of Pittsburgh, School of Medicine, Department of Medicine. The University of Pittsburgh is a large academic research center with a population of approximately 4000 medical researchers spread over twelve institutions. The Department of Medicine, a department within the University of Pittsburgh School of Medicine, employs approximately 564 research-physician scientists (i.e., MD, PhD, PharmD), who actively work within a specific field of study (i.e., Basic, Clinical Trial, Clinical Outcomes, and Clinical Translational).

3.3 BIG SCIENCE

In 2005, The National Institutes of Health (NIH) decided to reinvent itself by the creation of a very large grant program called the Clinical and Translational Science Award (CTSA) programs, which has a focus on “Big Science” projects. This term is built upon the work of both Derek De Solla Price (1963) and Alvin M Weinberg (1967) and is used

to describe the large-scale approaches needed to solve the complex modern problems of medicine.

The NIH has determined that their new focus is on translational projects, i.e., studies that bridge the laboratory and the physician's office, also referred to as "Big Science" projects (Littman et al., 2007; Zerhouni, 2005a; Beaver, 2001). This direction, new for the NIH, is designed to truly transform human health (Zerhouni, 2005a). Many disciplines within the academic institution are focused on this new roadmap; this dissertation asks the question "What can the academic medical library do to help achieve this goal?"

Traditionally, university medical institutions employ Library Information Science (LIS) professionals to gather, organize, and classify documents (e.g., abstracts, books, articles) , which are used by medical researchers as an information source for their publications or grants. Can we reverse this process, use a Library Information Science method that categorizes documents, and direct this process to the classification of medical researchers into distinct subgroups (i.e., also referred to as the medical researcher's "field of study") with the purpose of assisting "Big Science" projects?

3.4 THE CLASSIFICATION PROCEDURE

The development of an automatic classification system is proposed, based on an ontological method used within Library Science (Miller, 2001, p.245-246), known as the Resource Description Framework, i.e., "an infrastructure that enables the encoding,

exchange and reuse of structured metadata,” which was also adopted as part of the Dublin Core Initiative. Morville (2005, p. 131) describes the Resource Description Framework as a W3C (World Wide Web Consortium) standard used for describing and exchanging metadata, which is used as a general method of modeling information by the utilization of subject-predicate-object expressions, commonly called triples, [see Figure 3-1].

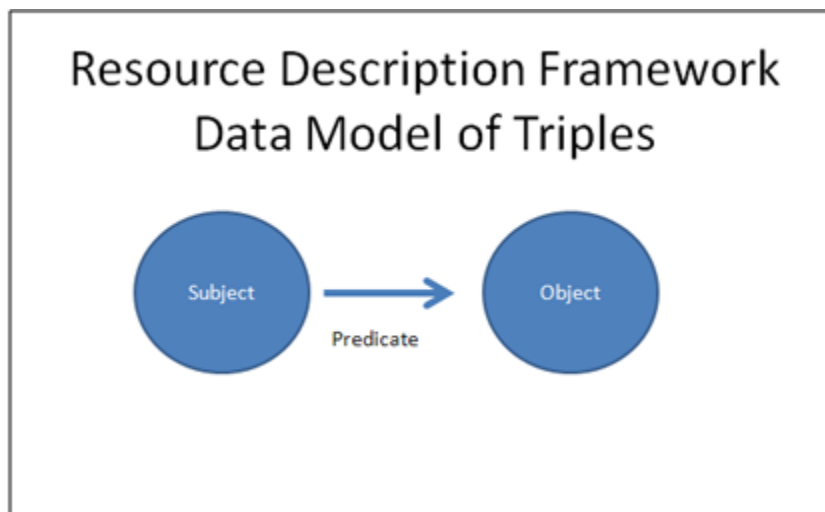


Figure 3-1 Subject-predicate-object expressions

This procedure extracts words and/or phrases from abstracts located in the public databases of PubMed®, <http://www.pubmed.gov>, and the Computer Retrieval of Information on Scientific Projects (CRISP), <http://crisp.cit.nih.gov/>). The extracted words and/or phrases are used to develop associated subject-predicate-object triples, which are then used to classify the abstract into a “field of study.” Later, using an algorithm on the academic medical researcher’s abstract classifications, the procedure classifies the

researchers into their field of study and a level of activity (i.e., the medical researchers are classified as either very active or active).

This classification procedure searches medical researcher-specific documents (i.e., abstracts) and then extracts metadata (i.e., “structured data about data” (Miller, 2001, p. 245)), which is used in a Classification Procedure to classify the research professional into an appropriate field of study and their level of activity. The metadata consists of words or phrase that are “equivalent to” or “related to” the specific field of study of the investigator, [see Figure 3-2]. The count of each medical researcher’s publications is used to categorize the level of activity of each medical researcher.

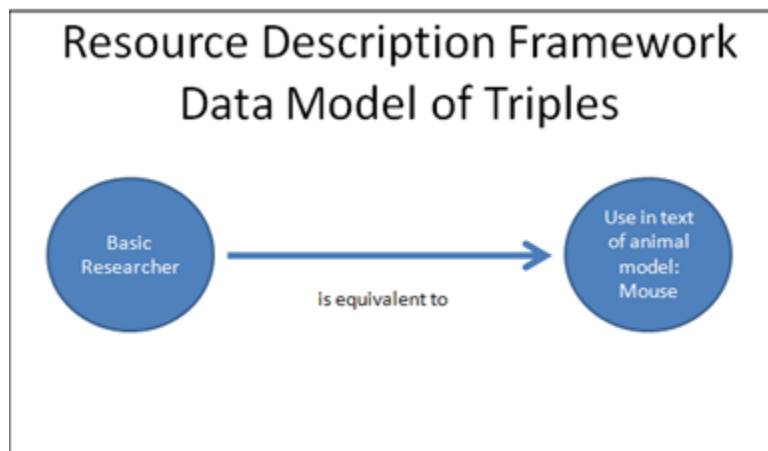


Figure 3-2 Metadata consists of words or phrase that are “equivalent to”

The focus of this dissertation is to determine whether an automatic classification procedure can match the classification ability of a domain expert (i.e., a senior research investigator whose institutional position has been to guide, assist and allocate resources

to a wide range of senior and junior investigators within a specific division of that institution). The use of a domain expert is common in bioinformatics reliability studies (Hripcsak & Heitzan, 2002; Rosenbloom et al., 2008) and this dissertation proposes that an automatic classification system, which can mimic the classification skills of the domain expert, would provide a novel tool to the academic medical library and a needed service to the academic medical institution.

The Classification Procedure identifies subgroups within the institutional research community and identifies the investigator activity levels within that subgroup. This classification procedure setup requires two team members from the institution, the domain expert and a LIS professional (i.e., the informationist).

3.5 TEAM MEMBERS: THE INFORMATIONIST AND THE DOMAIN EXPERT

This initiation procedure for the classification system depends on two individuals with different skill sets. The first is a library-trained individual, the informationist, also referred to as an information officer, an interdisciplinary professional with skill sets derived from library science and informatics. The role of the informationist in providing highly specialized services in the clinical setting is well known (Rankin et al., 2008, p. 194). The informationist has been described (Detlefsen, 2002), concerns about their role (Sathe et al., 2007, p.270), and suggestions of the informationist's supportive role within team environments (Hersh, 2002; Detlefsen, 2004) have been proposed. The classification procedure developed within this dissertation is designed to be part of the toolkit of the informationist, extending the role of the informationist from the clinical to

the research setting. The procedure requires knowledge of electronic database querying techniques, utilization of web services, and an understanding of programming languages to extract and parse text-based information.

The informationist, using various electronic database-querying methods, discovers all published documents of a specific medical researcher. The publication search results determine the most active members within each field of study using a simple counting mechanism. This dissertation assumes that the publishing level of each medical researcher is directly proportional to the activity level of that medical researcher.

Price and Beaver (1966, p. 1011) refer to the highly productive publishing research members as the “power group” and suggest that they control the administration of research funds, as well as allocation of laboratory space. Drenth (1998, p. 220) in his article, noted an increase in senior authorship and speculated that this was caused by senior scientists becoming more active in initiating, mentoring, and guiding research. The implied importance of activity level classification is the ability to identify the research members, especially from among the senior members, who are in the best position to assist in control and allocation of resources for their subgroup.

An interesting question can be asked, “Can the fusion of library science and informatics techniques create a simple method of classifying medical researchers into their appropriate subgroups (i.e., field of study) and identify the most active members of that subgroup?”

The second individual required for the initial setup of the classification procedure is the domain expert. The domain expert is a senior medical researcher who has an administrative or supportive role over the medical research subgroups and understands the important characteristics of each subgroup (e.g., the domain expert could identify a geneticist as a basic scientist who tends to use one of two specific core laboratories located within the institutional domain).

The classification procedure requires the domain expert's knowledge of the formal and informal authority structure within the academic medical institution, especially the ability to understand the unspoken rules (Rosenberg, 1996) that forms around medical research. As an added benefit, the domain expert knowledge is "written down" transforming the classification process from "black box," i.e., anything that has mysterious or unknown internal functions or mechanisms (Merriam-Webster, 2008, black box)" to a 'white box" process, which provides an observable, detailed record of the classification process.

4.0 THE CLASSIFICATION PROCEDURE AS A RESOURCE

There have been formal discussions on how best to use the resources and services of the academic medical library (Hart et al., 2000; Weise, 2004), with very creative and unconventional discussions of how to attract patrons e.g., supply consumer services (Houlihan, 2005), and present interesting future visions of library services (Marcum, 2003). This dissertation presents a similar novel service, an automatic classification procedure, which can be used within an academic administrative environment to help with fiscal decisions for allocation of scarce resources. The development of an automatic classification procedure is useful only if it has real significance, and to do so, important questions must be answered.

- 1 How can automatic classification solve complex problems?
- 2 Why use a document based classification procedure?

4.1 HOW CAN AUTOMATIC CLASSIFICATION SOLVE COMPLEX PROBLEMS

As an indication of importance, the automatic classification procedure must have the potential to solve complex problems found within institutional academic “homes.” The

section entitled “*Solving the National Institutes of Health mentoring problem*” presents a complex problem that uses the classification procedure to provide a library-based solution to a Clinical and Translational Science Awards problem. The section entitled “*Laying the foundation for push technology*” proposes how the classification procedure can be used to provide novel new services to assist with the complex information needs within the medical research population.

4.1.1 Solving the National Institutes of Health mentoring problem: An example of the classification procedure creating a novel new service.

The University of Pittsburgh is in the process of identifying the basic, clinical trial, clinical outcomes, and clinical translational science groups (especially the clinical translational researcher) for inclusion into an 83.5 million dollar NIH Roadmap grant (Rossi, 2006; Whelan, 2007). The academic medical library, as an established resource provider, is well positioned to assist the academic medical research professionals with their information needs and should have a place at the decision making process of use of these funds. There are many professional organizations, even within the field of medicine (Ewigman, 2008), that are trying to be included in the decision making process. It is very competitive and to ensure a place at the grant table, the academic medical library must demonstrate that they can provide new services that add value, while at the same time being financially responsible.

The NIH is also concerned with teaching the next generation researchers and a major component of teaching is mentoring. Even though the NIH is moving in a new

direction, mentoring has always been a historical and ongoing concern of the NIH programs. Dr. Elias Zerhouni, the Director of the NIH, referenced mentoring, in 2005, as an important component of the NIH budget when he stated:

*“In an attempt to address these concerns, the NIH has funded facilities, resources, or both to bolster clinical and translational research, such as the General Clinical Research Centers, grants for individual or institutional training and **mentoring** [bold added], support for disease-specific centers, clinical-trial networks, biospecimen repositories, molecular-screening libraries, and more recently, loan-repayment programs designed to attract and retain scientists to this field. Currently, the NIH spends about 36 percent of its budget on clinical research and training activities. Yet, the concerns persist, and more must be done” (p.1621).*

Zerhouni (2005b, p.1356) again expressed his concerns in reference to mentoring when he said, *“...the exploding clinical services demands and shrinking financial margins at academic health centers have limited protected research time and curtailed the **mentoring** [bold added] of young investigators.”*

The focus on mentoring continues with the creation of the CTSA program, described by Dr. Susan Shurin, Deputy Director of the National Heart, Lung, and Blood Institute (NHLBI) in 2008 as a *“bold and unprecedented investment in the infrastructure of clinical research in the United States”* (p. 4). In this new program, Shurin (2008, p. 4) stresses mentoring as an important element within the new CTSA program that will be addressed.

*“To realize this vision, the NIH has created a research consortium of Clinical and Translational Science Awards (CTSAs) that will include institutions across the United States working as a consortium to bring new treatments to patients, develop innovative approaches to clinical and translational research, support training and **mentoring** [bold added], of investigators, extend clinical research into the community, create robust and interoperable research informatics, and develop interdisciplinary teams” (p 4).*

But, as reported in the Senate Appropriations Committee report accompanying the Fiscal Year 2007 budget, the first step of the CTSAs process has been largely focused on training, as reported by Morrison (2008, p.8), *“The [CTSA] initiative appears focused largely on the training of new clinical investigators and may result in a diminution of resources currently available to active clinical researchers.”*

Mentoring is considered an important part of academic research (Sambunjak, 2006) and is an important part of librarianship (Kwasik, 2006; Davidson, 2006). Mentors serve a variety of important roles (National Institutes of Health, 2008), but very few financial resources exist that can identify potential mentors within an academic medical institution. A method for determining a pool of potential mentors would be beneficial to the less productive faculty within the academic medical institution and provide a pool of talent that would fulfill a requirement of the CTSA grant model.

Mentoring is now in danger of becoming sidelined due to underfunding and micromanagement of time utilization of investigators. Nevertheless, mentoring plays an important role in the career of the research investigators, especially as they compete for scarce resources during the grant writing process (Cole, 2006). The NIH emphasizes (Bhattacharjee, 2007) the importance of mentoring by including mentoring, (under the

heading of “*consulting with colleagues and graduates students*” - US Department of Health & Human Services, 2008, Question 16, 3rd paragraph), as percent effort on research grants.

Mentoring is a difficult process and dependent on individual investment of effort. English (2003) describes mentoring as a, “*one-on-one process of selecting and grooming promising candidates from the apprenticeship pool*” (p. 10). One senior scientist at the University of Pittsburgh described the mentoring process as a vital but extensively time-consuming process in which a single senior faculty member could properly mentor two, but at most, four trainees or junior faculty members a year (confidential personal communication, September 2008).

There is universal agreement that mentoring is an important and necessary step that assists both junior and senior investigator investigators as they move their careers forward. Unfortunately, there is no agreement on how to identify possible mentors for the large numbers of medical researchers, such as those found at a large medical university (i.e., the approximately 500 clinical translational researchers that exist at the University of Pittsburgh).

Identifying and then asking successful research investigators (both senior and junior) to spend time in mentoring is difficult. Compounding the problem is the need to identify specific mentors within the NIH targeted medical research subgroups. In a commentary (Woolf, 2008), the author states that translational research is a priority for

the NIH, with an expectation that 60 centers would be funded with a budget of \$500 million per year. The author explains translational research as a *“bench-to-bedside” enterprise, tying together knowledge from basic sciences with the intent of producing new drugs, devices, and treatment options for patients.*” Mentorship is considered very important within translational research, since this collaborative design has different rules and challenges (Poher et al., 2001).

Mentor identification, an important component of the CTSA grant model, can be considered a subset of a classification problem that divides medical researchers into appropriate subgroups and identifies their activity level. This dissertation suggests that the “very active” members of a subgroup could serve as a pool of mentors who could assist the remaining active researchers within their specific field of study.

Therefore, to accomplish the task of identifying targeted sub-groups within a NIH funded institute, this dissertation demonstrates a library-based method that automatically classifies University of Pittsburgh, School of Medicine, Department of Medicine, medical researchers into domain-specified subgroups and determines their level of activity.

4.1.2 Laying the foundation for library based push technology: An example of the classification system as an information filter.

Support for medical researchers and helping them with their careers is an important part of the mission. Price and Beaver (1966) note that *“there exists a core of extremely active researchers and around them there is a floating population of people who appear to collaborate with them in one or two multiple-authorship papers and then disappear not to be heard from again”* (p. 1004). Rosenberg (1999) speaks of the *“progressive, dangerous decline in the number of physician-scientists”* (p. 331) and calls for a national database that would identify and track physician-scientists who enter and leave the research environment (p. 332). Identification of those medical researchers who emerge in a research subgroup, transfer to another subgroup, or leave the field all together would be an important resource of information for the mentors and for the library. The academic medical library, by following the publication activity of the research author, can determine if that researcher is still working within a specific field, has left that field, or has left the research world completely. The use of a document-author based classification procedure would provide “level of activity and field of study” identification of researchers (both junior and senior) as they flow through or out of the various research subgroups. In other words, the medical researcher’s publication history can indicate if researchers are currently publishing research articles or if years have passed since their last article. A publically available database, along the lines of the Rosenberg model, populated with medical researcher information, could be created quickly.

A novel service that academic medical libraries could use to keep the medical researchers interested in their research careers would be to provide relevant information in a timely manner. Push technology is an example of an underused informatics tool that librarians (Clemmons & Clemmons, 2005; Gustitus, 1998) and others (Kendall & Kendall, 1999) have investigated, but not fully utilized because of its inability to live up to its stated potential (Lepori et al. 2002). Push technology was popular in the mid-1990s (Franklin & Zdonik, 1998) which promised to "push" news and other information to computer desktops with no user intervention. Push technology is currently being used at the National Library of Medicine in its Drug Literature Program, which transmits vast arrays of clinical, research, and toxicological drug data to a diversified group of individuals (Knoben, 2004). Knoben praises the use of "push technology," which the author suggests, " *allows timely transmittal of new medical information that can be customized to individual interests*" (p.172).

One of the major problems with push technology has been information overload (Edmunds & Morris, 2000), which refers to the tendency of technologies, such as push technology, to direct large amounts of possibly irrelevant information to the user. Manual methods of filtering the information flow and sending only relevant information to a user have been reported. Neill (1989) suggests that an "information analyst" could act as a "filter to identify quality research papers" for a user. The information analyst would search for information requested by users, develop a profile of their needs, and acting as a "human filter," be able to select relevant documents based on the information analyst's understanding of the users' information needs.

Unfortunately, these manual methods are restricted by the limited work-time available to information professionals, and consequently scaling to a large number of users is impractical. Automatic methods that filter out the irrelevant information and then develop information profiles for the user have been proposed; examples include a method by Hust (2005), which uses information derived from previous searches of the collaborative team and Klink (2004), which *“learns with the help of feedback information of all previous users (and also previous queries of the current user).”* Both methods require that baseline information be examined prior to the creation of an information profile and not be used to create an initial default information profile. Lam et al. (1996) stresses the importance of an automatic mechanism to continuously add information to the information profile and to follow changing user interests for personalized information filtering.

An interesting approach to information filtering was suggested by Laine-Cruzel et al. (1996), in which the authors defined a user profile that consisted of “stable information, related to a particular person rather than to a particular search” and “variable information, related to a specific search” to personalize the search and limit the information to what is relevant for the user. The library tool of citation analysis has been used to find “stable information” about a particular person. Kademani & Kalyane (1998) developed portraits of individual scientists using citation analysis, which examined characteristics of the scientist, such as his or her subject specialization. The “stable information” derived from citation analysis, can act as a filter, creating an information profile that restricts information flow to only relevant information. This relevant information could then be transmitted to the user by the use of “push technology.”

This dissertation suggests that the automatic classification procedure could also provide this “user profile” through its classification of the medical researcher into a specific subgroup. This classification procedure, which uses the researchers’ own publications to categorize, should provide the “stable information” suggested by Laine-Cruzel et al. The classification procedure insures that only information relevant (e.g., new journals, new services, links to multimedia information) to the medical researcher’s subgroup would be targeted for transmission to that medical researcher subgroup (i.e. field of study).

Therefore, the answer to the question “*Why should the library consider automatic classification important?*” is because an author-document based classification procedure would solve complex problems and would demonstrate that the academic medical library is an excellent knowledge steward since it can target specific library resources to specific subgroups.

4.2 WHY USE A DOCUMENT-BASED CLASSIFICATION PROCEDURE?

Interviews, surveys and questionnaires have been used as library classification tools for some time (Hoskisson, 1997; Hallmark & Lembo, 2003). In an article by Ried et al. (2006), a questionnaire survey was developed that asked researchers to self-classify themselves into one of four research categories, based on previous research experience. The categories presented by the authors were, “*non-participants (little or no*

previous experience in research); participants (as part of a research team); managers/trainers (either leading research, or in formal training to do so); and academics (with, or leading toward, a doctorate) “(p .2). An article by Porter (2001) describing the results of the 1993 National Study of Postsecondary Faculty survey states, “*The NSOPF survey asked faculty to choose their principal fields of teaching and research from a detailed list of academic disciplines and major fields of study*” (p.180).

However, interviews and surveys raise major concerns; an example in an interview-based classification system is the tendency of the research user to focus on current events rather than generalizing over the long term (MacLean et al. 1998, p. 146). Another concern is response rate; an example presented by Ried et al. (2006), noted that, “*All 229 members were posted a survey questionnaire,*” with only 89 members responding, which is “*representative of only a proportion of members' background, skills and needs.* “

However, the two biggest concerns with interviews and surveys are time and cost. As reported by Crowley et al. (2002) “*A single interview can last several hours* “(p. 207). In a similar manner, a survey needs, at the minimum, to be designed, sent out, and filled out (i.e., even in the electronic version, someone must “click a link” and then take the time to fill out the pages). As early as 1979, Shosteck and Fairweather estimated the data gathering expenses of a properly-run survey (initial design, setup, follow-up, contacting, resending, etc) at \$63.00 per subject. Even if cost were not a factor, acquiring updated information would still limit the usefulness of surveys and interviews; for example, how many surveys or interviews would individuals agree to participate in before they stopped responding?

Self-identification also has problems; what researchers think about their “field of study” and what the institution considers the individuals’ “field of study” may not be the same. In other words, researchers may not classify themselves in a similar way. Carpenter (2007) provided an example of why this could occur in an article, describing a conversation with Dr. Anthony Hayward, clinical research director for the NIH’s National Center for Research Resources. Dr. Hayward acknowledged that a career path in translational research was very risky and that the main reason for this is limited funding, suggesting that translational research has not been well funded in the past. Other concerns are the collaborative nature of translational research, which affects publications rights, intellectual property rights, and grants, since the norm is to credit only one primary investigator. Therefore, a translational researcher may not wish to self-select the translational research field.

The last concern is scalability, or the ability to ask multiple questions. A medical researcher has limited time and may object to multiple surveys and questions. However, it is a simple matter to data-mine a document for information, add or modify the questions, and then re-mine the document again.

This suggests that asking medical researchers for “field of study” information multiple times over the year (since their research focus, as Price and Beaver suggest, may change) is time-consuming, costly, limiting, and may not provide the answers needed. To simply gather this information, the academic institution commonly uses a domain expert to identify the “field of study” of the medical researcher, because of the domain expert’s long experience. Therefore, an automatic classification procedure that uses the knowledge of the domain expert for input can overcome these limitations.

So, an answer to the question, “Why *use a document-based classification procedure?*,” is that a document-based categorization procedure can provide the ability to re-analyze, rapidly respond to new questions, and remove scalability concerns. An automatic classification program can analyze 400, or 4,000 or 40,000 researchers with no additional cost.

5.0 LITERATURE REVIEW SUMMARY:

This literature review is divided into three parts: (1) personalization systems as a filter for groups; (2) a brief look at a document-based data mining method, i.e., citation analysis, used in a novel way, as a classification method to differentiate populations into subgroups; and (3) an examination of the information-seeking behavior of specific academic medical researcher subgroups.

The following databases were searched for supporting information: Library and Information Science Abstracts (LISA)®, Library Literature & Information Science®, Library, Information Science & Technology Abstracts (LISTA)®, MEDLINE® and Cumulative Index to Nursing and Allied Health Literature (CINAHL)®.

5.1 PERSONALIZATION SYSTEMS AS A FILTER FOR GROUPS

A number of different methods have explored user personalization, which attempts to filter information to the individual and the group. One example is conceptual clustering, which develops collections of similar documents, using an internal document function. Authors (Godoy & Amandi, 2006; Michalski & Stepp, 1983) have suggested various types of document clustering algorithms that are based on the content of the document. The advantage of this method to connect with user preferences and interests is the

ability to develop categories without a priori knowledge of the type of categories needed. Document clustering is another similar method (Leuski, 2001), which is based on the concept that closely associated documents tend to be relevant to the same user request.

Recommender systems are another method of personalization. Herlocker et al. (2004) described the recommender system as using the opinions of a community of users to help members of that community narrow their choices from a broad range of options. As described by Perugini et al. (2004), these systems reduce information overloads by extracting a subset of items from a universal set based on user preference. The recommender systems are designed to connect people together and exist within user networks. These systems can use content-based filtering, which recommend items that the user has liked in the past. Alternatively, these systems can use collaborative filtering, which recommends items based on the preferences of similar users. Herlocker et al. suggests that collaborative filtering is successful (p.6) and has a number of useful algorithms, but does admit that identifying the best algorithm for a domain is difficult.

Besides filtering content, the recommender systems can suggest other areas of interest that may be useful to the end-user. The recommender systems can also connect groups of individuals with similar interests together by using a “similarity” function, which may be based on closeness, distance or nearest neighbor algorithms.

There are difficulties associated with recommender systems, with issues such as sparsity, (i.e., very few similar items or users), over-specialization, (i.e., a focus on a

small select group of items), as well as privacy and trust issues. Adomavicius and Tuzhilin (2005) suggest improvements to recommender systems, such as support for multi-criteria ratings and an incorporation of contextual information into the recommendation process that they feel will enhance recommender systems in the future.

Faloutsos and Oard (1995) describe another filtering method -- natural language processing (NLP) -- which attempts to match queries with the semantic content of documents. Automatic summarization, described by Jones (2007), is a subtask of NLP that attempts to extract the main topic of the document. Summarization could be used to group together documents that are of a particular concern to a specific group.

Automatic indexing has a long history (Stevens, 1970) and involves the selection of words or phrases to identify content with documents. Hoyle (1973) describes a method of filtering information using an automatic indexing process that assigns documents into nine categories. Hoyle's method uses Bayes's Theorem to determine the probability of a category based on a word that occurs in the document.

5.2 AN EXAMINATION OF A DOCUMENT CLASSIFICATION SYSTEM: CITATION ANALYSIS

Classification, defined in the Merriam-Webster Online Dictionary (2008) as the "systematic arrangement in groups or categories according to established criteria," can

be the basis of a system used for studying the specific needs of the professional in a variety of fields. A classification system can be complex; MacLean et al. (1998) investigated a method of identifying research priorities, with an interest towards allocating research funds, in the public sector by examining the dynamics of a value-added chain that consisted of services, consultancies, businesses, interest groups, and government authorities. One of the tools developed to examine this relationship was an environmentally-derived research user classification system, based on an interview process, which mapped users into 20 sectors (e.g., oil and gas, construction, universities, energy, etc).

Classification system can be as simple as a binary system; Will (2006) wanted to determine if a new journal would potentially have enough new authors and a sufficient audience for publication. The author used data mining to classify a prospective audience of 1600 authors into those who might be interested and those who might not be interested in writing for the new journal. The conclusion suggested that data mining might be an easy method of analyzing the research behavior of the authors (p. 1049).

Within library science, individual professional categories have been recognized in the past as having an effect on the professional's use of library resources. An example is the work by Powell (2002), which observes the impact of the professional association (ALA, ASIST, MLA, SLA) membership on member's library usage, "ASIST members read an average of three research journals, followed in decreasing order by members of MLA, SLA, and ALA" (p. 69). Librarians (Leckie et al. 1996, p. 162) recognized that commercial interests have been developing services, which focused on the different

information needs of scientists and scholars (i.e., the professional category of the academic specialist).

Information science, especially in the areas of artificial intelligence and document retrieval, has recognized that professional user classification is important. An article by Chu et al. (1999) describes three user classification systems, *“non-domain-specific which characterize users by the extent of their knowledge, domain-specific which use stereotypes to describe general groups of users, and multidimensional approaches that combine non-domain and domain-specific techniques”* (p. 103). Petrelli et al. (2004) continue this observation when they note that, *“We met journalists, analysts, translators, and librarians and discovered they differ in search experience, language knowledge, and final goal”* (p. 928) and suggest that different professionals can be classified into different *“search classes with different user needs.”*

An interesting document-based method that uses text-mining techniques to create a variety of categories according to some established criteria, such as, user groups, investigator profiling, social structures, or structured teams within an academic setting, is citation analysis. Text mining as suggested by Kostoff (2002, p. 2789) refers to the approaches used to analyze and extract useful information from text. An approach used to develop research user profiling was developed by Kostoff (2001), using a blend of text mining and bibliometric analysis. Text mining, and its associated concepts, i.e., citation mining and citation pattern analysis, has been used within library science, especially as a method of analyzing electronically-stored unstructured text data (March, 2008), legacy data (Tan, 2007) and to determine hierarchy (e.g., teams, groups, social structure) within a population.

The literature reports on numerous uses of article citation pattern analysis to differentiate authors into groups. There are a number of reported concerns associated with citation analysis as a journal evaluation tool (Garfield, 1979a; Phelan, 1999; Porter, 1977; Garfield, 1979b; MacRoberts & MacRoberts, 1989; Kostoff, 1998), but citation mining and the analysis of citation patterns do not seem to have the same concerns.

Social structure as a foundation and explanation for group dynamics has been discussed. There have been reports of citation patterns used to illuminate social structure or highlight collaborative teams within larger research populations. White, Wellman and Nazer (2004) suggest a relationship (i.e., a social network of citations) between co-citation and acquaintance, with an emphasis on intellectual rather than social ties. The authors examined the personal relationships and communication behavior among an international group of 16 researchers. The researchers were drawn from seven disciplines and were focused on studying human development. The authors were able to show that for this group, citation patterns do have a tendency to reflect social structure. As the length of membership within the group increased, their articles demonstrated an increase in citation rates with each other. In reference to collaborative work, White *et al* (2004) report, “Interciters tend to be working on a joint project or reading each other’s work or coauthoring something.” This suggests that information that is useful to one member of a collaborative team may be useful to other members of that same team, i.e., homogeneity within a subgroup is valuable since a single resource, like a specific journal, can have high value among multiple people.

Within communities such as a university setting, collaborative teams rapidly form and dissipate depending on their funding base or collaborative needs. It is difficult to

find these teams, but an article by Ichise, Takeda and Ueyama (2006) proposes that research communities built around a particular researcher can be found using three relationships, co-authorship, citation, and author citation. Rapid identification of teams within large communities (e.g., a large academic university or large pharmaceutical company) could be used as a method of determining what newly formed teams are “in the pipeline” and will eventually have information needs.

A document’s citation patterns, used as a method of separating academic levels within an academic university structure, have been proposed. Ventura and Mombrú (2006) concluded that citation profiles of full and associate professors differed. In their article, the population was similar in age, but the full professors showed differences in the number of papers per year and their citations counts. The number of citations per paper was not influenced by multi-authorship or by internationalization of the papers by that particular author. This article suggests that the citation profile combined with number of papers per author, per production year, may be usefully incorporated into the development of the policy used in the promotion of associate professors, and as a method of ranking the associate professors to determine who would receive promotion to full professor. Using this method, the authors found a statistically significant difference between full professors and associate professors. In another instance, citation patterns have been used to differentiate levels of undergraduate students. Magrill and St. Clair (1990) looked at differences in citation behavior by course level and in different disciplines. Their paper suggests that citation counts increased from the sophomore and juniors to senior level. The authors collected 1775 undergraduate term paper bibliographies or footnotes derived from their regular course assignments in

selected departments of four different institutions. Students in the humanities used books rather than journals, and scientists used more journals than books. Students in the sciences also cited more journal articles and from a wider range of years than the other students. Tailoring transmitted information to team members would be a useful feature and the use of the above method to aid in the identification of professional levels within a team, would be one way of providing appropriate information to the right individual.

The characterization of an article's citation pattern for basic science researchers within journals in specific specialties has been proposed. Adusumilli et al. (2005) determined that basic science research publications within the larger universe of United States general surgical journals had significant citation frequencies. The authors noted that these general surgical journals were also important since they formed a "transitional bridge" between laboratory and clinical research. Within the general surgical journals, the authors found that basic science research publication "is cited 32 times (range 1–141, median 11)." The ability to identify a pool of information that could be tailored to a specific group, i.e., general surgeons, should be beneficial and provide an easy method of characterizing group prior to the first meeting.

Classification of users by team roles (i.e., using team structure as categories) has also been investigated. Teams tend to have "star" members who have a definite standing within the research group and this would suggest citation patterns as an identifying tool to spotlight specific authors. Hill and Provost (2003) write that citation patterns, along with a referee's personal background, could identify authors who submit their work to a double-blind review process for scholarly research articles. The

assumption is that because of the “star” author’s intense focus in his or her very narrow area he or she can be recognized by their very tight citation patterns.

Cox et al. (1994) looked at citation patterns in the specific area of anxiety disorders research. They found that there was limited evidence of citation use across journal and author disciplines, and specifically determined that psychiatric journal publications rarely cited psychological journal publications.

Herubel and Buchanan (1994) suggest that citation pattern analysis is useful for determining the characteristics of a discipline’s literature. In the example of the social sciences, variables such as gender, institutional affiliation, productivity ranking, obsolescence, and format can be derived by citation analysis. Additional variables such as subdisciplinarity and emerging discrete research fields may also be derived. Citation analysis may be able to obtain the specific characteristics of the literature that is of interest to a particular collaborative team, thus creating a profile for that team.

Citation mining can provide insight into the type of publication a clinical trials author may be interested in examining. Peritz (1994) looked at clinical trials publications and determined that: citing authors partially preferred large studies to smaller ones, focused their citations on publications that presented the minority view of the research, and tended to cite papers published in “high circulation journals.” Peritz suggested that citation analysis could be used to determine the current or historical interests of the authors of articles on clinical trials.

Joswick and Stierman (1997) performed a comparison of the most frequently used journals by faculty and students of Western Illinois University. The authors

proposed that the referenced lists were very dissimilar between these two groups, and concluded that within one academic library, citation patterns differ markedly between user groups.

Turati, Usai, and Ravagnani (1998) used citation analysis as an instrument to explore research frameworks, within the Academic International Research Projects (AIRP), attempting to find commonality between researchers working in Europe and the United States. The project champion used citation analysis to find team members who were more likely to have the commitment to “bridge” the differences between the two cultures. The conclusion of the authors is that, “the intensity of relations among authors is represented by the number of same references in two scholars’ bibliographies” (p. 195). Here the authors divide the team into those who are “collaborative” and those who are “not collaborative” in nature (Turati et al., 1998).

The above authors suggest that publications, using an established library method (i.e., citation analysis), can be used to classify professionals. This dissertation suggests that other parts of the publication (i.e., abstracts) are just as useful in classifying professionals.

5.3 AN EXAMINATION OF THE DIFFERENT INFORMATION SEEKING BEHAVIOR OF ACADEMIC MEDICAL RESEARCHER SUBGROUPS

Sung et al. (2003) suggests that the NIH considers the clinical research process as fragmented and not functioning as a cohesive whole. The implication is that the clinical

researchers are clamoring for resources without adequate oversight of the resources. Demonstrating that library resources can be targeted to specific sub-groups would suggest that the library could aid allocation of resources on an as-strictly-needed basis, and that this would lead to a greater understanding of the resource needs of the whole group, and could lead to additional library funding opportunities.

A method that differentiates medical researchers into their specific sub-groups and identifies their most active members could provide library administrators with a “champion” who could guide expensive print and electronic resources to the appropriate research sub-groups. The very active members, who use the most resources, would be in a position to offer guidance on the prudent allocation of resources within the subgroups.

An example of a controllable resource is electronic journals access control. Those electronic journals that are common to all subgroups can be shared, with library costs sustained by the entire group, and electronic journals specific to a subgroup could be limited, with the library costs covered by those within the specific subgroup. If the specific subgroup does not consider the library resources valuable, then reallocation of library resources to another subgroup could follow.

Leckie et al. (1996) examined the information seeking behavior of professionals (i.e., engineers, health care professionals and lawyers) and determined that each group has specific library needs. Engineers considered journal literature irrelevant (p.165). Nurses rarely used the library (p. 169) and family physicians valued informal consultations with colleagues over journals and textbooks (p. 170). The library needs of

lawyers depended on the specific field of law that was practiced (p.173). This literature review suggests, as was the case with the lawyers, that the medical research subgroups also have different library needs and wants depending on their specific field of study.

Basic research, as defined by Calvert (2006, p. 199) is research that is directed towards acquiring “new” knowledge rather than trying to find a practical application from “old” knowledge. The School of Pharmacy (UCSF, 2003, heading: basic science research) located at the University of California suggests that the basic researcher is interested in more fundamental aspects of the life process, such as observing how cells operate. This suggests that information-seeking behavior for the basic researcher might center on problem-solving at the most focused level with the need for very specific, narrowly focused library resources. One possible practical implication is that literature searches for basic research subgroups would need to focus on the most current published articles with delivery measured in hours rather than days.

Calvert and Martin (2001) maintain that basic research is difficult to define but suggest that basic researchers share common qualities that differentiate them from other types of researchers. Talja and Maula (2003) report that differences exist in the way basic researchers use electronic journals when they state “*teaching versus research orientation, local versus international research orientation and basic research versus applied or action research orientation are factors that are likely to influence information-seeking strategies and e-journal use*” (p. 677). They are suggesting that “basic research” and “applied or action research” are different in the way information strategies are used. This implies that the information-seeking behavior of basic

researchers differ from clinical and translational researchers.

Other literature sources echo the same results by noting that basic researcher behavior appears to be directed primarily towards the basic science journals. In an article by Hurd, Bleicic, and Vishwanatham (1999), they report,

“The citation data show that the journals used most frequently by molecular biologists are basic science journals rather than medical titles, as classified by Ulrich’s” (p. 41).

An article by Brennan et al. (2002) again provides evidence that basic researcher information-seeking behavior is different, when they noted that the basic researcher used one type of database, in this case Web of Science®, rather than the discipline-specific resources such as Geo-Ref®.

Shine (1998) makes a case for the clinician as researcher since, *“only the well-trained physician scientist can thoroughly understand, interpret, and properly care for human subjects during studies that involve an intervention”* (p. 1442). The School of Pharmacy (UCSF, 2003, heading: clinical science research) from the University of California describes clinical researchers as medical clinicians who primarily conduct research on drug effects and other types of human interventions. This may suggest that information-seeking behavior in this group focuses on clinical medicine and outcomes research publications, as an approach to design novel methods of disease management.

Marriott (2002) describes the beneficial effects that electronic clinical journals have had on patient care. This would suggest that electronic journals were being discussed, probably during morning rounds, used in the routine care of patients, and would be expected to have influenced a number of clinical healthcare providers. The clinical researcher is also involved in the routine care of the patient (Snyderman, 2004) with the literature strongly suggesting that electronic clinical journals, used in evidence based medicine (Gralla, 1999; Guyatt, 2004), are useful for patient care and play a valued role in research studies that involve intervention and outcomes research. Even the method of delivery is important (i.e., the use of the electronic clinical journals, as opposed to the printed journal), for the busy clinical researcher, which is a relatively recent event as noted by Eysenbach (2002),

“Scientific communication and scholarly publishing are in transition. The age of printed publications as primary means to communicate research results is ending, being replaced by the era of electronic publishing (also known as e-publishing). This form of publishing has far-reaching consequences not only for how scientists distribute, access, process and digest information but also for how research itself is done and will be evaluated” (p. 499).

Andrews et al. (April 2005) reported that clinically based rural practitioners, as a subgroup, used more print sources rather than online sources. This may suggest that non-university based clinical researchers, who also are primary care practitioners, may have a tendency to use electronic journals less frequently than print journals. Korjonen-Close (June 2005), in a survey of clinical researchers, determined that their members felt that medical libraries were not providing them with the necessary information they needed and were requesting electronic resources, such as access to databases, online

journals and other health websites, rather than print material. The clinical researchers in this study also felt that the library websites were valuable but especially wanted access to high quality, evidence-based clinical and resource information. These observations imply that the quality of the electronic journals is important to the clinical researcher. If the quality journal is not available in electronic form, then the researcher will seek out the print form. So, the information-seeking behavior of the clinical researcher is influenced by their perception of what they believe the quality level of medical library electronic journals to be.

Translational research, as the Ontario Neurotrauma Foundation (2006) describes it is, “*the process of applying ideas, insights and discoveries generated through basic scientific inquiry to the treatment or prevention of disease or injury.*” (para. Translational research). The editorial by Pardridge (2003) further defines translational sciences as bridging the distance from the Petri dish to people. Mao (2002) discusses the information required to do this in an article, which stresses the need for bringing together both the basic and clinical aspects of pain research. This suggests that information-seeking behavior of a translational researcher is a collaborative venture between the “bench” and the “bedside” with the need to access publications of both types. The information needs of the translational researcher are based on very broad areas of information, often obtained from different fields of science, that require the integration of detailed clinical information about the patient collected over time.

The concept of translational research is relatively new (Pardridge, 2003) and discussions of the style of the translational researcher are now being debated (Zerhouni, 2006). Ioannidis (2004) observed the importance of translational science,

suggesting that translational research is a “mature” form of research derived from a fusion of basic and clinical research styles. This suggests that information and information-seeking behavior by the translational researcher contain elements derived from both basic and clinical styles of research.

An editorial by Pardridge (2003) comments that the translational researcher is concerned with what happens in the laboratory and tries to link that finding with outcome research derived from a clinical trial. Translational research by nature is multidisciplinary; examples exist within the literature (Mao, 2002) proposing that the translational researcher may have difficulty bridging the gap between the laboratory results and their application to some patient-related problem. This implies that the information-seeking behavior of the translational researcher is a collaborative or team-based venture between the “bench” and the “bedside,” with information needs based on very broad areas of information, often obtained from different fields of science. Research on manually-identified translational researchers suggest that their requests have little need for speedy retrieval, but do require a broad range of information from both bench type studies (e.g., animal studies) and clinical studies (e.g., outcomes research). This probably reflects the concern of translational researchers with the safety of human subjects over utilization of a new procedure.

Other authors also consider translational research rooted in collaborative behavior, as observed by Sonnenwald and Pierce (2000),

“In many dynamic work situations, no single individual can acquire the varied and often rapidly expanding information needed for success. Individuals must work

together to collect, analyze, synthesize and disseminate information throughout the work process” (p.461).

This suggests that the translational researchers need information from many sources, and that it would be expected that their information-seeking behavior would demonstrate a team-based approach.

This dissertation is attempting to identify and analyze subgroups within the population (both potential and actual) of the academic medical library. The identification and analysis of the user subpopulation within library populations is not new. Patron analyses exist in the library and are referred to as “customer intelligence.” Decker and Höppner (2006, p. 504) focus on the needs of the potential users of libraries and challenge the library to rethink their services. Decker and Höppner suggest the use of decision support systems can answer the question of “*Who are the customers and what needs and preferences do they have?*” (p. 507)

Classification of the actual library user or potential library user within the library population would provide insight into the type of library services that might be expected. An example would be a strategic decision involving the budget for space allocation for archived print journals within individual departments of a university. A classification system that identifies the types of user, the number of users within that subpopulation, and their archival needs would provide justification for the addition or removal of archive space from the resources of the library. As mentioned by Kavulya (2004, p. 118), “libraries are under pressure to justify their existence,” implying that a constant re-evaluation of library services is required.

5.4 SUMMARY OF THE LITERATURE

Citation analysis, which is a type of document-author classification, has been used to determine the makeup of the various professional populations that use the library.

Citation analysis is a method of examining citation links within a scholarly population; it is inexpensive, and relatively easy to perform.

Evidence found in the literature advocates that basic, clinical, and translational researchers act and work differently from each other. The basic scientist is described as being focused on a particular “new” problem, with their electronic journals needs narrowed to specific journals titles (Aerts et al, 1999, p.12). The clinical researchers, described as researcher-physicians, appear to be as concerned with their research as with their patients’ outcomes. Their electronic journal needs are focused on outcomes and intervention research, which appear to have a broader range than the electronic journal areas suggested for the basic researcher. A study by Andrews et al. (2005) suggested that within the clinical researcher field, variation exists between the non-academic and academic researcher. The translational researcher is a relatively new entrant into the research field. The literature suggests that their electronic journals needs are broad and involve reading journals that are outside their specific area of expertise, but within their research area.

Different research subgroups cite different types of publications, suggesting that the editorial makeup of the journal or the abstract may differ among the research subgroups. An example may be basic researchers, who cite only from single specific sources. This may imply that the language of the article or the abstract might reflect their narrow interest and contain a specific subset of words or phrases that may not exist in other research subgroups.

The literature review also demonstrates that user population analysis, a process sometimes referred to as customer intelligence, is already an established process within the library. The ability to use decision support tools to determine user needs provides the library with a mechanism to improve upon its services.

This literature review suggests that document based classification systems (e.g., clustering, indexing) have been discussed and methods proposed for use. Additionally, the literature review suggests that the library understands that specific subgroup publication requirements exist within the professional library population.

6.0 CONCEPTS

The document-author classification procedure is a new library tool that could be offered by the academic medical library to the academic institution, designed to classify the medical researcher population into “level of activity and field of study” subgroups. The total library population is defined as the potential as well as actual users of the medical library. The purpose of the classification procedure is to target library services and resources, both old and new, to specific “level of activity and field of study” subgroups within the medical researcher community, maximizing resource use and minimizing resource waste.

The document-author classification procedure locates abstracts of the “first and last author” in “PubMed®” and “CRISP” and downloads the abstracts to a local database, then combines abstract title and abstract body to create an “Abstract-Title” variable. Words are parsed, extracted, and tabulated from the abstract-title variable, the words from the “highly productive members” from each field of study are examined, and distinct “word/phase identifiers” particular to that field of study are determined.

The word/phrase identifiers become part of a “subject-predicate-object expression,” which acts as a “grouping mechanism” to classify the abstracts-title

variables into an “abstract field of study.” The abstract field of study is summed through an algorithm and used to assign the medical researcher to a particular “level of activity and field of study.”

6.1 FIRST AND LAST AUTHORS

Literature suggests (Shapiro, 1994, Buehring, 2007. 460; Drenth, 1998, p. 220) that within medical research, the first and last authors provided substantial leadership and guidance to a publication; in fact, Buehring observed that the final author on many occasions provided guidance to the first author, i.e., acting as mentor, (p. 460, para 2). This dissertation proposes that the first and last authors (i.e., also known as medical researchers) have a high interest in the publication content, which when analyzed can serve as an indicator of their “field of study.”

This study does not include in the classification procedure the additional co-authors that exist in the PubMed® record. Therefore, a researcher who publishes extensively as a non-first or non-last author would not show up in this study. The impact of other co-authors on the classification procedure will be analyzed in a future study.

6.2 PUBMED®

The National Library of Medicine, a NIH institution, has made available a searchable database of citations from Medline®. This electronic government-sponsored database is found electronically at <http://www.ncbi.nlm.nih.gov/pubmed/>. The PubMed® data is widely used as an archive for medical research articles, containing over 18 million abstracts for literature published over the last 58 years. For the purpose of this dissertation, the assumption is made that information about the majority of Federal government-sponsored research articles are archived at this site and the academic medical researcher would normally want to have the abstracts for their own work in this site.

6.3 CRISP

The Computer Retrieval of Information on Scientific Projects (CRISP) is a government database of federally-funded biomedical research projects, which have been conducted at universities and other research institutions. The database, located at <http://crisp.cit.nih.gov/>, contains project-specific information including institution and primary investigator name. For the purpose of this dissertation the assumption is made that the majority of federal government-sponsored grant abstracts are archived at this site.

6.4 ABSTRACT-TITLE

The abstract-title convention consists of two parts: (1) the abstract body and (2) the abstract title extracted from the published article. The title and the abstract body are concatenated and inserted into a single variable field, thus the combined fields can be considered one analyzable unit. Cohen and Hunter (2005, p. 19) use a similar method in the discussion of techniques used with natural language processing and systems biology.

A question that can be asked is “Why use the abstracts, rather than the full articles, keywords, or mesh terms to identify the subgroups,” since authors such as Pitkin (1987) and other have found the quality of the abstract to summarize the article sometimes to be lacking. The “Instructions to the author” section of medical journals, such as the Journal of the American Medical Association (JAMA), stress that major findings, results or outcomes of the submitted paper should be contained within the abstract (JAMA instructions for authors, 2008). This dissertation is using the abstracts as a surrogate, to determine if written material from medical researchers can be used to separate the medical researcher into their “field of study.” A further study may examine if the addition of the text of the journal article itself increases the accuracy of the classification procedure. This dissertation suggests, but does not prove, that the journal article is designed to speak to the world, while the abstract speaks to the researcher’s peers.

A second question that can be asked is, “Why concatenate the title and abstract into one variable?” This dissertation proposes that the title field and the abstract field

acting together *emphasize* what is important to the medical researcher and therefore can be used to determine the “field of study” for the medical researcher. Here we are building on the suggestion of Derek De Solla Price (1963) when he writes that we publish for small groups, “communicating person to person, instead of paper to paper.” (p. 91)

6.5 HIGHLY PRODUCTIVE MEMBERS

Rothman (1979) and Drenth (1998) propose that productive senior members of a profession exert professional control over the less-senior members of their research cohort. The underlying mechanism for this control is the productivity of the highly-productive members who use their experience to guide less-productive members. The highly-productive members within the medical research domain are expected to occupy the upper end of a publication-based power curve. The written publications of a few of the highly productive members are associated with each research subgroup, and this dissertation proposes that word/phrase identifiers associated with the highly-productive members can differentiate a domain-specific population into research subgroups.

6.6 WORD/PHRASES IDENTIFIERS

An interesting example of how words can be used to differentiate subgroups is found on the internet (McConchie, 2002; Campbell & Plumb, 2003) at

<http://popvssoda.com:2998/>. This website, influenced by an article written by Von Schneidemesser (1996), lists the geographical location of individuals together with their word preference for ordering carbonated beverages (i.e., using the word “soda” or “pop”). This website provides a java-based, interactive nonrandomized collection site that visually promotes the use of the words “soda” and “pop” to identify a person’s geographical location. The suggestion from this website is that people tend to use language associated within their social area.

In a like manner, Shultz (1996), Green (1986), Nwogu (1997), and Rothman (1979) have reported that professionals create their own languages (i.e., jargon) and do this for many reasons. Jargon “conveys information concepts and requests succinctly (Shultz, p.45),” “label conceptually complex material (Green, p. 365),” and even “mystify” those outside their area of specialization (Rothman, p. 499). This dissertation extracts words and phrases, a type of domain-specific jargon that appears to be distinct to specific research subgroups, for the purpose of differentiating the subgroups. As an example, in this study, the phrase “GROUPED INTO” was found to be primarily used with outcomes researchers and “CLINICAL PERSPECTIVES” was found to be primarily used with trial researchers.

Thus, this dissertation proposes that distinct words and/or phrases within the abstract can provide important clues as to the subgroup identity of the medical researcher.

6.7 SUBJECT-PREDICATE-OBJECT EXPRESSIONS

This dissertation focuses on the development of a simple procedure that classifies medical researchers by the words and/or phrases found in PubMed® or CRISP Abstracts. This tool borrows concepts from an ontological tool known as the Resource Description Framework. An ontology, as used in library science, is defined by Gruber (1993, p. 1) as “an explicit specification of a conceptualization.” This dissertation uses a method of conceptualization referred to as the Resource Description Framework, a general method of modeling information using Subject-Predicate-Object expressions, commonly called triples. There are two relationships, based on the triplet, that form the basis for the procedure created: (1) “is equivalent to;” and (2) “is related to.” Using these expressions, this study has created a procedure that equates a “field of study” within a specific domain (i.e., the fields of study within the Department of Medicine), with a specific set of words and/or phrases. Additionally, there are words and/or phrases, when weighted appropriately, would suggest a relationship between the medical investigator and his or her field of study.

6.8 GROUPING MECHANISM

User self-classification within a domain, i.e., the tendency of users within a large group to break into subgroups, has both a biological and organizational basis. Large groups, from a biological point of view, appear to have a limit to their social size with Dunbar (1993) suggesting an upper limit of 150 members. Dunbar also notes that language may

function as the cohesive glue that holds the members together. This strongly suggests the existence of a lower biological size limit on the number of professionals that can occupy one “field of study” within a domain. Knowing the number of professionals within a domain (i.e., Department of Medicine) one may be able to predict the number of professional subgroups prior to meeting with the domain expert (e.g., with 564 professionals within the Department of Medicine, one may expect no less than four subgroups)

The tendency of large groups to break into subgroups has been reported by other authors. Schein (1993) proposed that “*organizations of all sizes will show a greater tendency to break down into subunits of various sorts based on technology, products, markets, geographies, occupational communities, and other factors not yet known*”(p. 41). Schein implied that subunits within the organizations develop their own subcultures with “different languages and different assumptions about reality” (p. 41).

Schein describes integration within an organization as primarily a problem in meshing subcultures; these subcultures begin to define themselves and set their psychological boundaries by developing their own language. Schein proposes that, “*we will need technologies and mechanisms that make it possible for people to discover that they use language differently*” (p. 43). The proposed classification procedure uses these languages differences to separate the subgroups.

It is interesting to note that Thomas Kuhn (1962) in “*The Structure of Scientific Revolution*” argues that scientists work in a series of paradigms, which reflect a predominant way of thinking and talking about a particular topic confined to a specific

professional group (e.g., transplant surgeons writing about a suture method). This may suggest that the subgroups (e.g., Pulmonary Transplant, Liver Transplant), when they band together into larger professional community structures, echo this language separation. Transplant surgeons as a group write in a similar manner to each other and write somewhat differently from cardiologists.

The proposed classification procedure depends on the natural tendency that people have to self-group with specific languages within subgroups acting as a boundary, separating one “field of study” from another.

6.9 FIELD OF STUDY

The field of study variable consists of “Basic,” “Clinical Outcomes,” “Clinical Trials,” and “Translational” subgroups. Researchers such as Ioannidis (2004), Crowley (2003), and others have used these terms to describe a medical researcher’s area of specialization.

This dissertation proposes that a classification procedure can analyze the concatenated abstract-title and determine the “field of study” associated with that particular abstract-title. An algorithm then takes the series of abstract-title “field of studies” associated with a particular medical researcher and categorizes that medical researcher into a specific “field of study” and later also adds the “level of activity.”

6.10 LEVEL OF ACTIVITY AND FIELD OF STUDY

The “level of activity” is added to the “field of study” to create a new outcome measure because of the temporal component of medical research. If the researcher has written extensively in a field of study over a period of years (i.e., in this study defined as 12 research or grant papers or more over a 6.5 year period), then one may expect that the researcher will continue to write research papers in that field of study. If the researcher has not written extensively in a field of study, then the future direction of that researcher is unknown. So testing for both types of classification (i.e., deciding if the researcher is an active basic researcher or if the researcher is a very active basic researcher) may be useful to predict those who may continue to publish in that field of study.

Publications as a measure of research productivity have been discussed; Cockburn and Henderson (1996) describe a highly productive scientist involved in pharmaceutical research as one who publishes extensively, e.g., more than 20 papers per year. Publication counts as a means of advancing university researcher status and attaining tenure have been discussed by Vucovich et al. (2008) and others.

This dissertation proposes that publications counts, confined to when the medical researcher is first or last author, can act as a qualitative measure of the level of research activity for that particular medical researcher. The publication counts, referred to as “level of activity,” can then be combined with a medical researcher’s “field of study,” to classify medical researchers into their subgroups.

7.0 METHODOLOGY

This dissertation tests the hypothesis that an electronic method of categorizing academic medical abstracts and researchers into their “level of activity” and “field of study” performs as well as a domain expert’s categorization method. This dissertation’s focus is to see if an electronic categorization method can match a human-based method.

The domain expert is a Professor of Medicine & Pharmacology and the Director of the Center for Clinical Pharmacology. The domain expert has experience as a senior faculty mentor, has served as the director of the former General Clinical Research Center (GCRC), and is now involved with the Clinical and Translational Science Institute at the University of Pittsburgh.

One of the common institutional methods of allocating university resources to research members is to ask a senior faculty member to provide a recommendation of the needs of the medical researchers within their division. The senior faculty member reviews research member’s *curriculum vitae*, publication lists, awarded grants, and other publically available sources and provides a recommendation of the type and amount of university resources that should be allocated to the medical researcher.

7.1 STUDY DESIGN

The study design is a retrospective, cross-sectional, inter-rater agreement study, designed to compare two classification methods (i.e., machine and human). The study population consists of University of Pittsburgh, School of Medicine, Department of Medicine (DOM) professionals who (1) have published at least one article listed in PubMed® as first or last author and/or (2) are the primary investigator for at least one grant listed in CRISP. The study population contains a potential list of 564 professionals, and of these, 305 professionals meet one or more inclusion criteria.

The text from the Department of Medicine electronic phone book was extracted, parsed and the first and last names of the professionals as well as their degrees, (i.e., MD., PhD, SCD, or PharmD), downloaded into a secure access database.

The information source consists of abstracts from the PubMed® and CRISP databases associated with the Department of Medicine study population, to be sampled over a 6.5 year period (2002- July 2008).

7.2 FLOWCHART

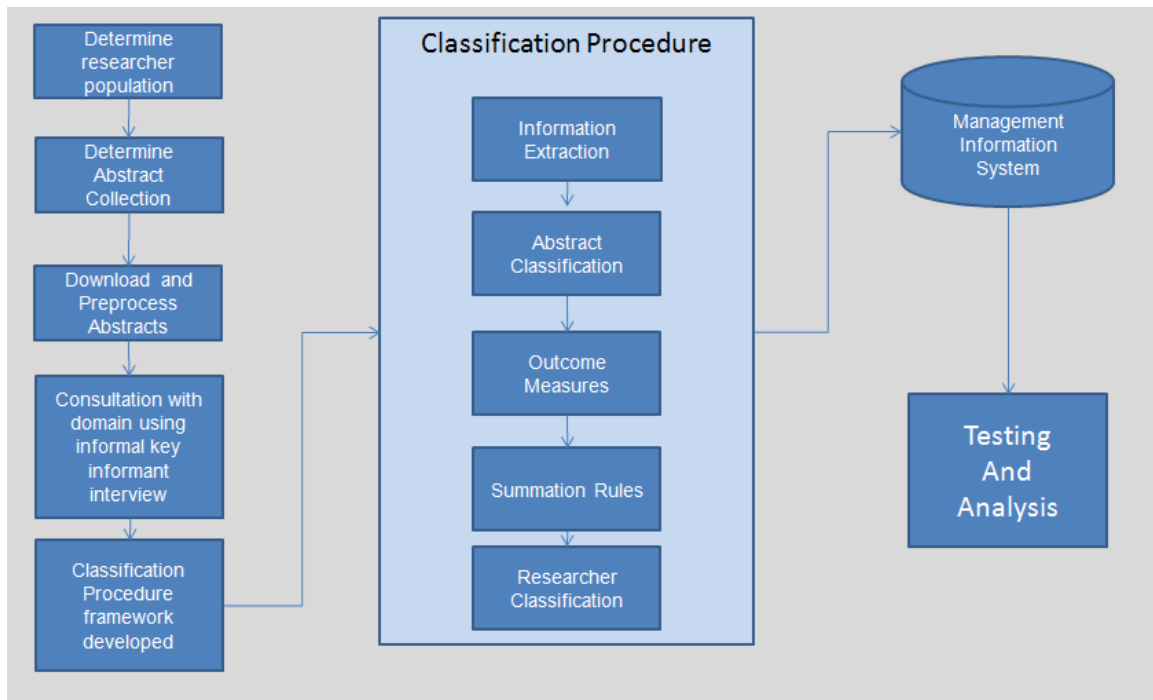


Figure 7-1 Classification flowchart

A medical researcher population that meets the inclusion criteria is determined. The location of the abstracts (i.e., PubMed®, CRISP) related to the targeted researchers is identified. The abstracts are downloaded and pre-processed.

An informal interview (2-3 hours) with the domain expert produced the information used in the classification procedure framework. During the interview, the domain expert provided a description of each researcher category. This description was used to find unique words or phrases within the abstracts that were used to classify the research groups (e.g., Basic researchers tend to work with an animal model, such as

“zebra fish”; Clinical Trial researchers tend to work on “phase 2 trials” with “controls”). The domain expert also provided two examples of a very active Basic researcher, Clinical Outcomes researcher, and Clinical Trial researcher. The Translational researcher examples are not requested, since the Translational researcher will be a blend of basic and one or both of the clinical researcher categories.

The example abstracts, identified by the domain expert, were parsed and examined to find words or phrases that appear uniquely in that category. The classification procedure, using a variation of rule-based text mining method (Cohen & Hunter, 2008), then searched for the words or phrases within the study abstracts. The words or phrases are passed through the “equivalent to” and “related to” functions described in section 7.7, which were used to create the abstract classification outcome measure. The outcome measures were stored in a management information system, tested against the domain expert’s categories, and the abstract results were then analyzed.

The classified abstracts, using the summation rules in section 7.9, were then used to categorize the medical researchers, into their “field of study.” Finally, the “level of activity” rule, which is presented in section 7.8.2, was used to further subgroup the medical researchers into their “level of activity” and “field of study.” The outcome measures were stored in a management information system, tested against the domain expert’s findings, and the researcher classification results were then analyzed.

7.3 INCLUSION CRITERIA

1. The medical researchers must be listed in the July 2008 Department of Medicine departmental phone listing.
2. The medical researchers must have University of Pittsburgh-affiliated publications or grants.
3. The medical researchers must have at least one PubMed® publication where they are first or last author and/or at least one grant in the CRISP database where they are primary investigator.
4. The medical researchers must have at least one publication or grant between the years 2002 and July 2008.

7.4 LIMITATIONS

The background of the informationist, a multidisciplinary profession, is important, since an understanding of the role of the medical researcher is needed. An informationist with a medical background should be able to duplicate the above findings; an informationist with an unrelated background may not.

The academic environment at the University of Pittsburgh, one of the top NIH funded institutions in the country, is one of constant grant competition. The classification procedure depends on the tendency for the very few “star” professionals

(highly published researchers with high grant funding) to selectively recruit, mentor, and train other, less senior, less funded professionals. As an example, a highly productive Basic researcher will select less senior researchers who are expected to work alongside the Basic “star” professional. Since the less senior researchers will probably use resources provided by the “star” professional, there is an expectation that the abstracts will contain standardized abbreviations for these resources. This institutional structure may not exist at other less grant-funded institutions.

The project is collecting historical information from PubMed® and CRISP and there is no guarantee that all abstracts or grants have been published online. There is no guarantee that all abstracts or grants specific to that author have been collected.

The classification tool only analyzes medical researchers who are first or last author on the abstract or grant. If the medical researcher does not appear as first or last author, the medical researcher is excluded from classification.

The classification program does not take into account any publications prior to employment at the University of Pittsburgh.

This classification program examined word patterns (i.e., CCP, RNA, CDAD, discharge) and no attempt is made to understand the meaning of the word or to use stemming. In other words, “mice” and “mouse” are considered separate patterns rather than singular and plural forms of the same object.

A single domain expert, an accepted practice, was used to identify the field of study for each abstract and medical researcher. Each domain expert provides his or her

own rules for classification; it is unknown if the results would vary with another domain expert.

The variables “very active” and “active” are a first pass at understanding how productivity affects the researcher’s classification process. Within the domain of this study, two or more publications (i.e., articles, grants or both) a year were considered difficult to produce by this domain expert. This may not be the opinion of other domain experts and may not be true in other research domains or professions.

7.5 ASSUMPTIONS

An assumption is that PUBMED® and CRISP abstracts are correctly coded and have correct author sequence.

Medical researcher information (i.e., names and titles) come from the current (2008) online Department of Medicine directory listing and no attempt has been made to determine if researchers have left or if new researchers have joined. In addition, name variations have not been identified or corrected. Because this dissertation used an online internal phone directory, correct spelling of the researcher’s name was assumed.

This study suggests that the first and last authors have the highest interest in the publication content. Therefore, this study assumes that the abstract content and the title closely reflect the first author and last author’s interests. A future study will examine this assumption on the other co-author classifications.

7.6 PUBMED® ABSTRACTS

A web service (i.e., eUtils.eUtilsService, http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_ms_help.html) was used to query the PubMed® database. The web query:

1. Search for the medical researchers in first (i.e.,[1AU]) or last position (i.e.,[LASTAU]);
2. Search only for those medical researchers affiliated with the University of Pittsburgh (i.e., AND (UNIVERSITY OF PITTSBURGH[AD])); and
3. Only accepted abstracts in PubMed® or CRISP between the years 2002 and 2008 (July) (i.e., (2002[EDAT]: 2008[EDAT])).

A PubMed® abstract was linked to the medical researchers only if those medical researchers were listed on the abstract as the first or last author. The only exception involves those abstracts that list a group as the last author. In this case, the next to the last author position was accepted as the last author. The date of inclusion in PubMed® was used as date of the abstract.

7.7 CRISP ABSTRACTS

All CRISP abstracts associated with the University of Pittsburgh, to include the primary investigator name and date of abstract, were downloaded and abstract information parsed to an MS Access database. A CRISP abstract was linked to the medical

researchers only if that medical researcher is the primary investigator. The CRISP start date was used as date of the CRISP abstract.

7.8 ABSTRACT-TITLE VARIABLE

The PubMed® and CRISP abstracts and title from each of the medical researchers were downloaded, concatenated, and inserted into an abstract-title variable. The abstract-title variable consists of the title of the abstract plus the abstract body as retrieved from PubMed® or CRISP. Author names were removed from the abstract and de-identified by associating each medical researcher name with a study number. The classification procedure analyzed the abstract-title variable content and determined the field of study for that abstract based on that variable.

The researcher classification procedure is a multistep process that categorizes the abstract-title variable into “field of study” selections and then inserts current time-period abstract-title classification (for research question 3 the abstracts counts are also added) into a summation algorithm, which then categorizes the medical researcher.

7.9 DOCUMENT-AUTHOR CLASSIFICATION PROCEDURE FRAMEWORK

The Document-Author classification procedure is a tool designed specifically for the librarian and follows the following framework.

1. The librarian acquires an electronic listing (e.g. electronic phone book, departmental list) of all domain specific professionals who have used or have the potential of using the academic medical library.
2. The librarian consults (informal key informant interview) with a domain expert to discuss the definition of each ‘field of study,’ see Figure 7-2. The informationist must pay careful attention to what activities separates each “field of study” from the others. For example, in the Department of Medicine, basic researchers tend to work in laboratories and typically experiment on research animals, while clinical researchers tend to work with hospital patients.

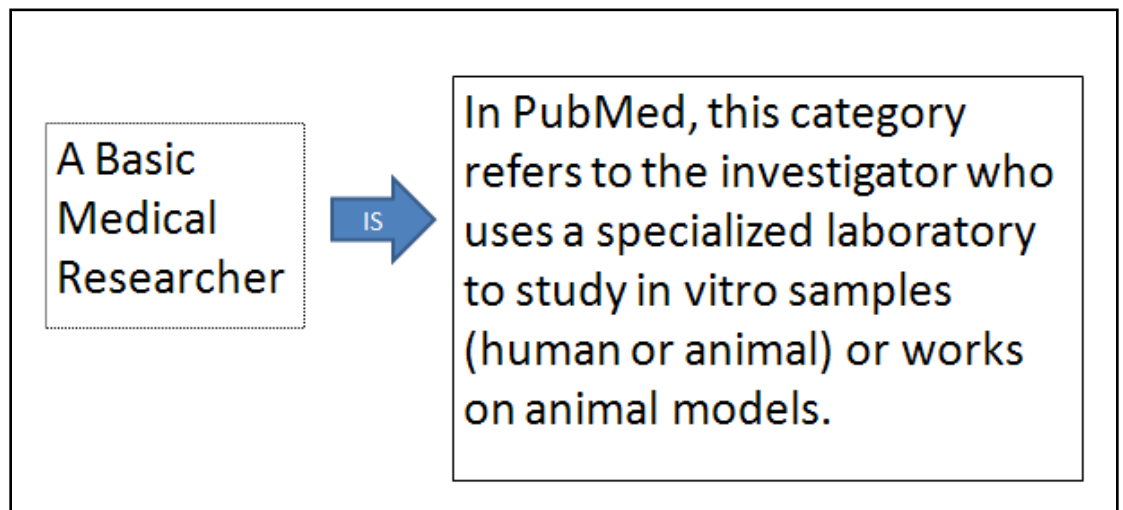


Figure 7-2 The informationist discusses the characteristics of each "field of study"

3. The informationist determines the location of the electronic resource used as source information. For this dissertation, the sources were PubMed® and CRISP Uniform Resource Locator (URL) addresses.

4. The informationist creates the query structure template used to access the electronic resource (e.g., for PubMed® the query template is last name, first initial, middle initial [1AU] OR last name, first initial, middle initial [LASTAU]) AND UNIVERSITY OF PITTSBURGH [AD], see Figure 7-3. The query string was inserted into the classification procedure.

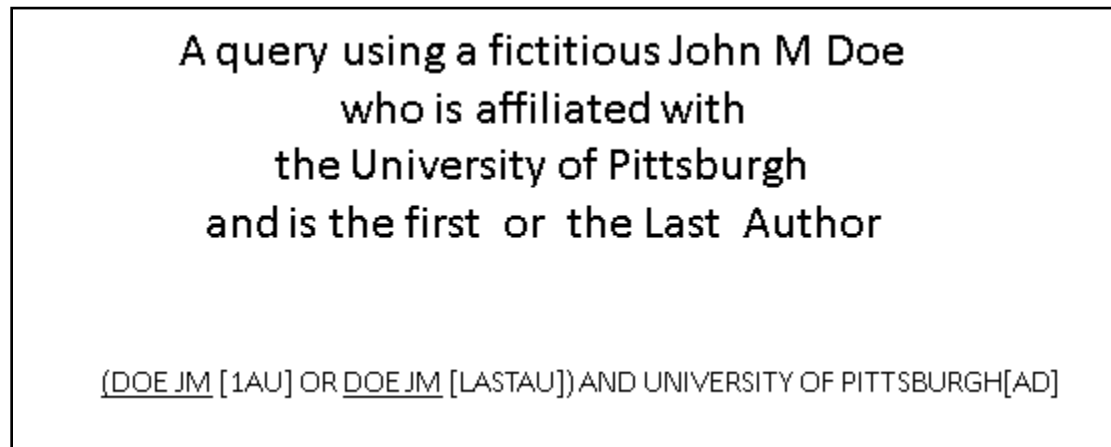


Figure 7-3 The query string

5. The raw abstracts and abstract title are concatenated and were downloaded into a secure access (Management Information System) database, see Figure 7.4.

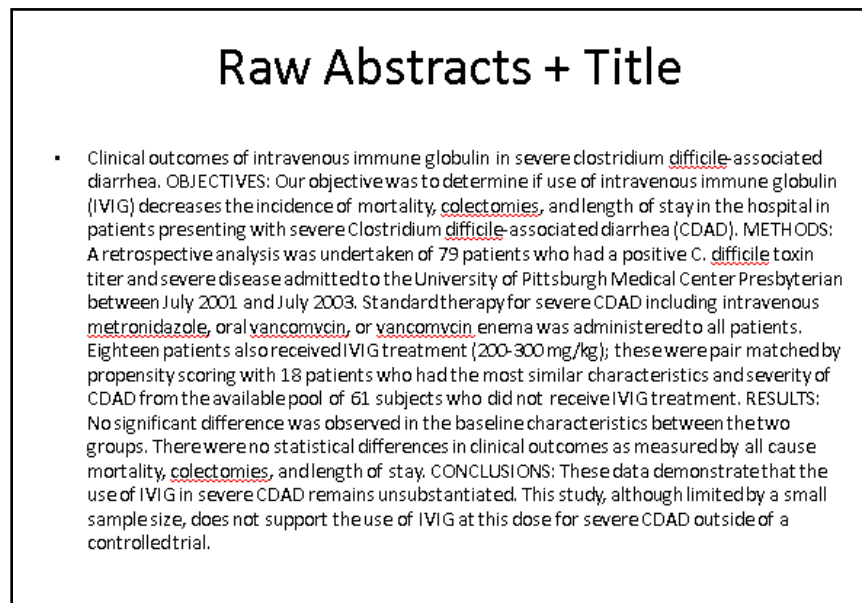


Figure 7-4 Raw PubMed® Abstract and Title concatenated into one Field

6. Two examples of very active research members in the Basic, Clinical Trial, and Clinical Outcome “field of study” groups were requested from the domain expert (total test member n = 6) for use in the pilot study. These examples were separated from the final study population and were only used in the initial definition phase. The reason for the very active research member request is the assumption that the less senior researchers will use whatever resources are made available to the very active researchers.
7. The Translational “field of study” group is treated differently because it will have characteristics of both basic and clinical. Translational “field of study,” classification

occurs when the abstract has characteristics of both the basic and clinical “field of study” subgroups.

8. The very active research test members (n=6) had a large number of abstracts. These research members were used for both pilot and final study to determine the word/phrase identifiers for each “field of study,” which were used to create the subject-predicate-object expressions.
9. For the study (pilot and final), the raw abstract + title variable is processed by removal of all punctuation [Figure 7-5].

Processed Abstracts + Title

- Clinical outcomes of intravenous immune globulin in severe Clostridium difficile associated diarrhea OBJECTIVES Our objective was to determine if use of intravenous immune globulin (IVIG) decreases the incidence of mortality colectomies and length of stay in the hospital in patients presenting with severe Clostridium difficile associated diarrhea CDAD METHODS A retrospective analysis was undertaken of 79 patients who had a positive C difficile toxin titer and severe disease admitted to the University of Pittsburgh Medical Center Presbyterian between July 2001 and July 2003 Standard therapy for severe CDAD including intravenous metronidazole oral vancomycin or vancomycin enema was administered to all patients Eighteen patients also received IVIG treatment 200 300 mg kg these were pair matched by propensity scoring with 18 patients who had the most similar characteristics and severity of CDAD from the available pool of 61 subjects who did not receive IVIG treatment RESULTS No significant difference was observed in the baseline characteristics between the two groups There were no statistical differences in clinical outcomes as measured by all cause mortality colectomies and length of stay CONCLUSIONS These data demonstrate that the use of IVIG in severe CDAD remains unsubstantiated This study although limited by a small sample size does not support the use of IVIG at this dose for severe CDAD outside of a controlled trial

Figure 7-5 The abstract + title field is processed

10. The downloaded abstract + title variable is processed (i.e., all punctuation removed, upper cased, parsed, only distinct words placed into the list) and the words are inserted into a table. In this process, a word is a continuous stream of letters or numbers surrounded by white space or beginning of file marker or end of file marker. A later study will examine the effect of hyphenated (e.g., time-period), possessive (e.g., patient's), and abbreviated (e.g., c. difficile) case words.

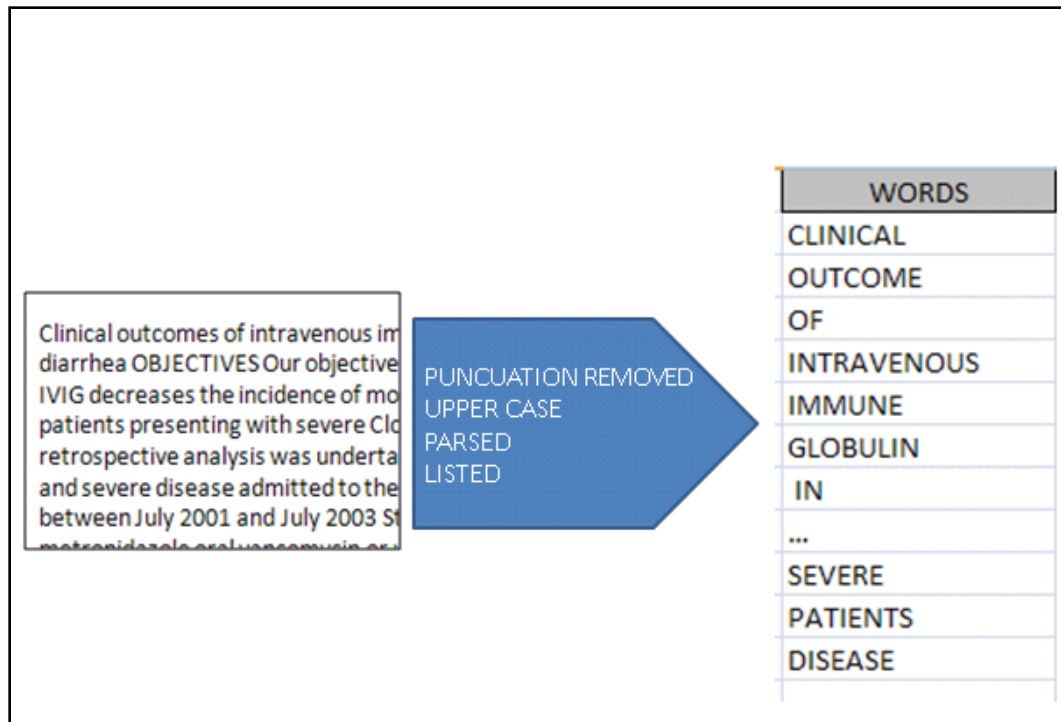


Figure 7-6 Abstract + Title processed into a table of words

11. An analysis of the “field of study” was performed on the initial 6 test member abstracts to compare the words from each of the very active “field of study” subgroups (i.e., basic, clinical outcome, clinical trial) to each other [Figure 7-7]. The lists of words in blue (taken from the basic abstracts) are compared to the list of words in gray (taken from the non-basic abstracts). The green words are the words that occur in the basic researcher subgroup of abstracts, but do not occur in the non-basic subgroup of researchers (i.e. clinical trial, clinical outcome).

Word or Phrases from the Target Group are compared to the Non-Target Group		
Target Group: Basic Researcher	Non-Target Groups	Only in the Target Group: Basic Researcher
WILL	WILL	
RESEARCH	RESEARCH	
DISEASE	DISEASE	
STUDY	STUDY	
GRANT	GRANT	
A	A	
IN	IN	
TEST	TEST	
THROUGH	THROUGH	
SPECIFIC	SPECIFIC	
CELLULAR	CELLULAR	
	INFLAMMATION	
	PATIENT	
	GROWTH	
	THERAPY	
	CELLULAR DISEASE	
CELLULAR MECHANISMS		CELLULAR MECHANISMS
MICE		MICE
RAT		RAT
CD4		CD4
CDNA		CDNA
CLAMPING TECHNIQUE		CLAMPING TECHNIQUE

Figure 7-7 Basic Researcher group compared to Non-Basic Researcher Group

12. The purpose is to find words that appear frequently in the specific “field of study” target subgroup and do not appear in the remaining “field of study” subgroups. An emphasis was placed on finding connected words/phrases (i.e., “clamping” expands as “clamping technique” in the word list derived from the abstract-title variable). A list that contains words unique to the target “field of study” was created [Figure 7-8].

Word or Phrases from the Target Group are not contained in the Non-Target Groups		
Target Group: Basic Researcher	Non-Target Groups	Only in the Target Group: Basic Researcher
CELLULAR MECHANISMS		CELLULAR MECHANISMS
MICE		MICE
RAT		RAT
CD4		CD4
CDNA		CDNA
CLAMPING TECHNIQUE		CLAMPING TECHNIQUE

Figure 7-8 Unique words in Target Group

13. The Resource Description Framework is well known within library science, used in standards for the creation and management of contents in digital libraries (Wu and Liu, 2001, p. 436). Using the Resource Description Framework as a guide, this framework creates a triple (i.e., Subject-Predicate- Object) that can link (i.e., make equivalent) a word to a “field of study.”
14. An equivalent triple would indicate that the abstract-title variable is equal to the selected field of study. As an example, a Basic “field of study” classification would be equivalent to an animal name (i.e., basic researcher - equals- mice). In other words, if the author has the word “mice” in their abstract-title variable, the researcher is considered a basic researcher in this domain [Figure 7-9]. In another domain, such as surgery, this may not be true.

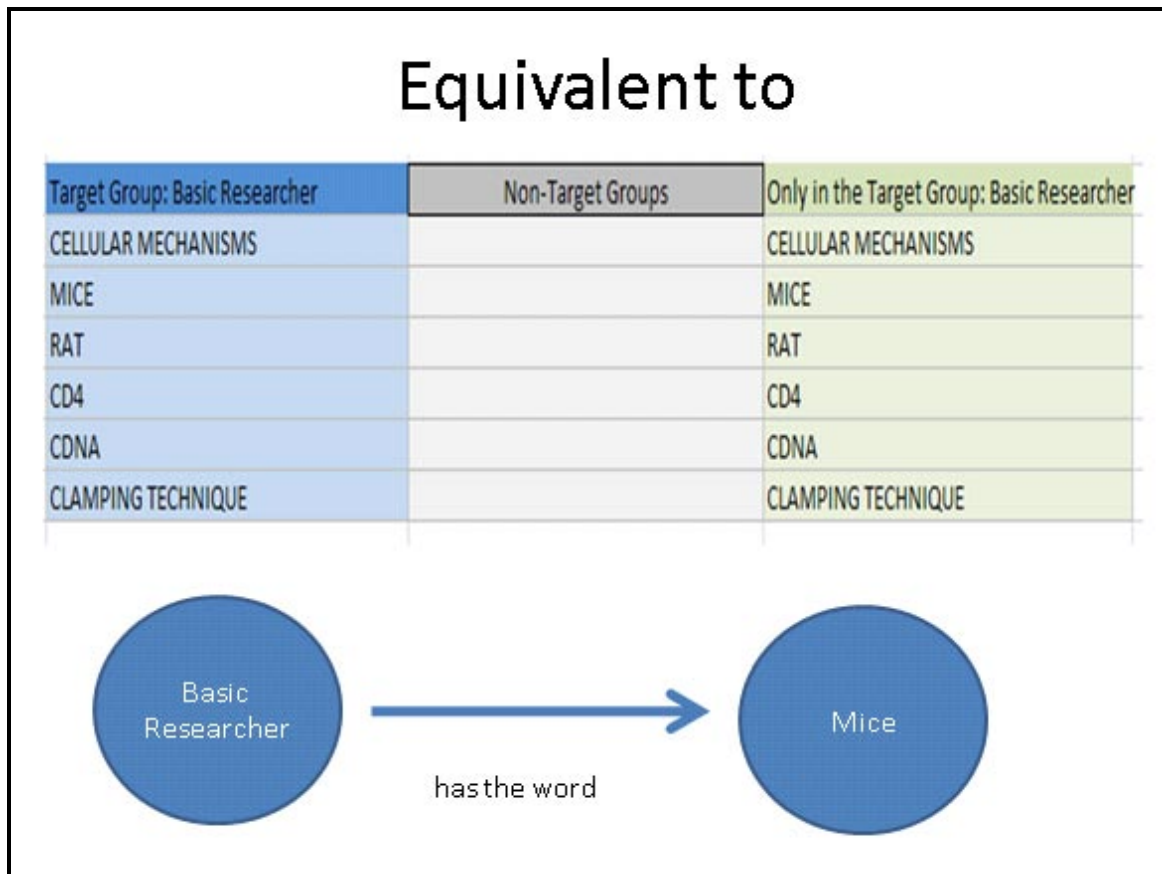


Figure 7-9 The word “mice” in the abstract-title variable is equivalent to a basic researcher

15. The test abstracts are classified using the “equivalent to” triplets. The pilot and final study results found that some, but not all abstracts are classified correctly.
16. To classify the remaining abstracts, another type of triplet is use. During the interview process with the domain expert, some characteristics of the “field of study” subgroups lead to an interest in certain words. As an example, clinical outcome researchers follow patients during the course of their stay in the hospital. After which, discharge summaries are examined. “Discharge” is considered an interesting word because it appears linked, in the opinion of the informationist, to the description of the clinical outcome researcher. Errors in the classification process are expected

to occur (i.e., some researcher may refer to discharge of fluids rather than discharge of the subject) but these errors are expected to be relatively few in this domain.

17. A list of interesting words and phrases belonging to the targeted subgroup (in this case Figure 7-10 show examples from the clinical outcome subgroup) are analyzed and found to occur in the other subgroups (i.e., these words occur in all groups).

Word or Phrases from the Target Group are contained in the Non-Target Groups		
Target Group: Clinical Outcomes	Non-Target Groups	Only in the Target Group: Clinical Outcomes
COMMUNITY	COMMUNITY	
COMPARE RATES	COMPARE RATES	
COMPARED WITH PATIENTS	COMPARED WITH PATIENTS	
COMPLEMENTARY METHOD	COMPLEMENTARY METHOD	
CONSECUTIVE PATIENTS	CONSECUTIVE PATIENTS	
CONTROL	CONTROL	
CONTROLS	CONTROLS	
CURRENT CLINICAL PRACTICE	CURRENT CLINICAL PRACTICE	
DEMOGRAPHICS	DEMOGRAPHICS	
DISCHARGE	DISCHARGE	
DISCHARGED	DISCHARGED	
END OF LIFE	END OF LIFE	

Figure 7-10 Interesting words from the clinical outcome "field of study"

18. A table with a field containing the counts of the interesting words for targeted "field of study" subgroup was created and compared to the other non-targeted "field of study" subgroups [Figure 7-11]. The purpose of the count field was to find words or phrases that occur frequently in the targeted "field of study" and less frequently in the non-targeted "field of study."

Word or Phrases Frequencies from the Target Group are compared to the Non-Target Group		
WORDS	Target Group: Clinical Outcomes	Non-Target Groups
INTERVIEW	4	1
SURVEY	6	2
DEMOGRAPHICS	4	2
DISCHARGE	4	1

Figure 7-11 Frequency count of interesting words

19. A related triple would indicate that the word/phrase has a high probability of being associated with the selected abstract “field of study,” i.e., the word “discharged,” appears in all “field of study” subgroups, but it appears more frequently in the clinical outcomes “field of study” subgroup in this domain.
20. A related triple, describes a word or phrase that when summed in the abstract, indicates that the abstract is a member of a particular subgroup, i.e., if the word “discharge” occurs more than two times in the abstract, this abstract was considered a “clinical outcomes” abstract [Figure 7-12]. The decision to select the value of 2 as the multiplier is arbitrary and based on a number that appears to work with all “related to” triplet words. Figure 7-13 shows selected sentences of a clinical

outcome “field of study” abstract-title variable that contains four occurrences of the word “discharge”

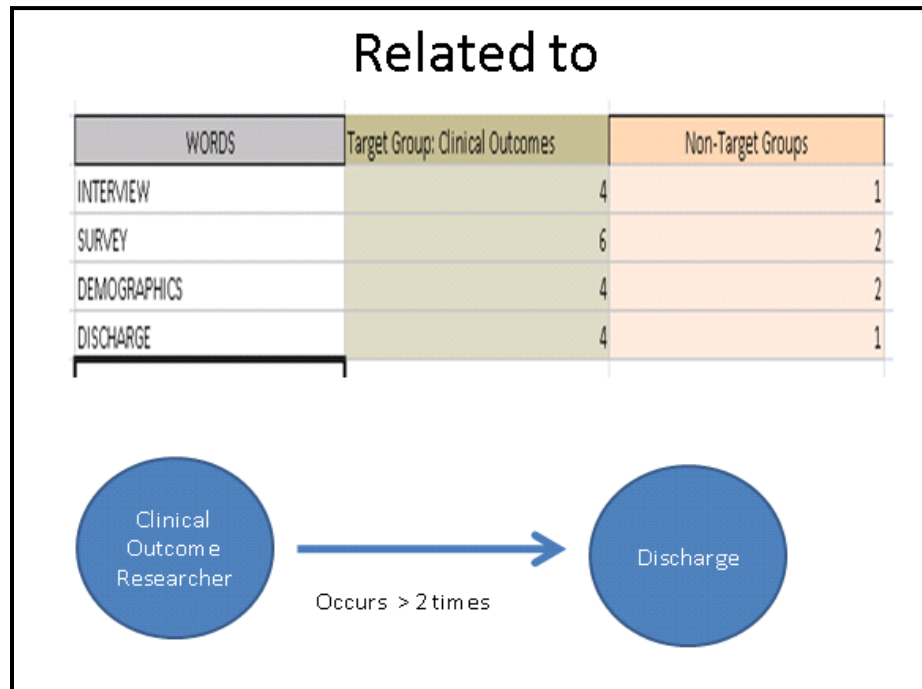


Figure 7-12 The word “discharge” is a "related to" triplet

- we performed a retrospective analysis of **discharge** data from 1995
- Outcomes included age, gender, race, **discharge** diagnoses (from ICD-9 codes) and **discharge** medications (from **discharge** summaries) in all patients.

Figure 7-13 The word "discharge" occurs four times in abstract number 454

21. The test population was analyzed and Kappa, a measure of inter-rater agreement, was used to determine success or failure of the above procedure.
22. The Translational "field of study" subgroup is a special case. An abstract is considered Translational only if the abstract-title variable contains basic and clinical classifications. In other words, the abstract-title has words that suggest a basic "field of study" and a clinical "field of study."
23. The abstracts are analyzed and classified by the "equivalent to" and "related to" triplets. In example 7-14, the word "enzyme" is found in the abstract-title variable. This abstract is considered (equivalent to) the basic "field of study".

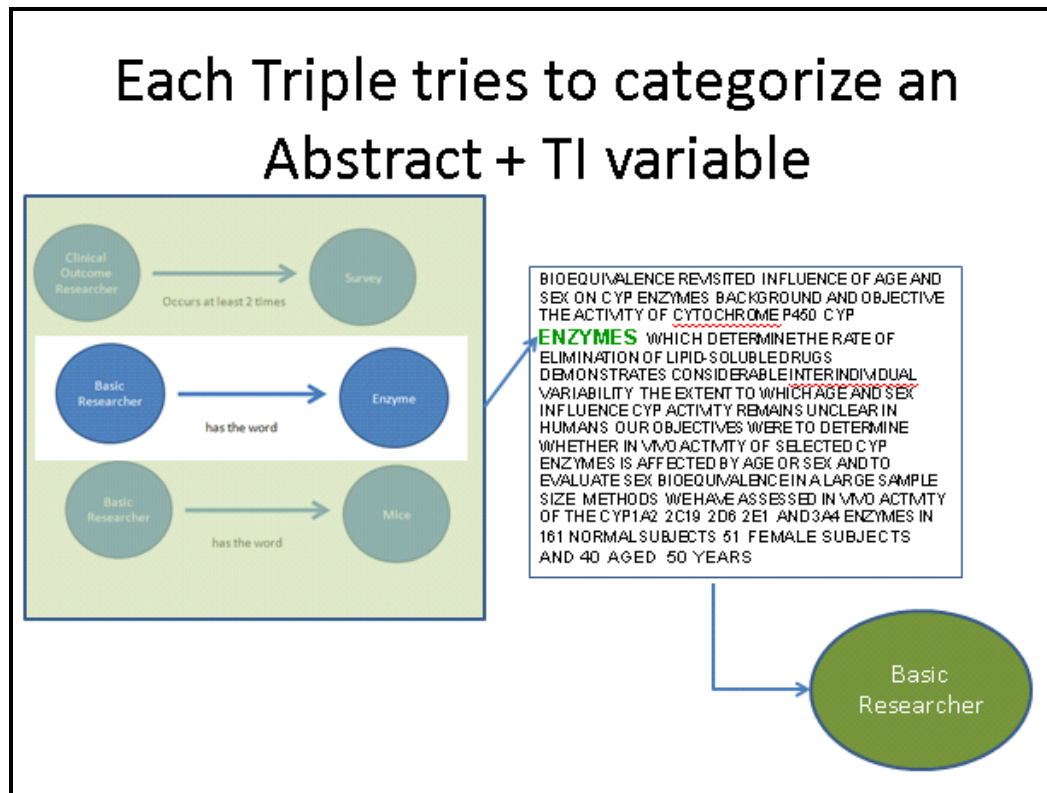


Figure 7-14 The triplets analyze each abstract

24. The medical researcher's "field of study" is determined by examining and summing the "current" abstracts "field of study" for each medical researcher.
25. For this dissertation, the domain expert considers the years 2006 to July 2008 as current years. This range was determined during the interview process; in a later study the effect of manipulating the current year range will be examined. If the medical researcher has abstracts from this period, those abstracts that fall within this period are considered, all others are discarded.
26. If the medical researcher has no abstracts within the 2006-2008 periods, the classification procedure took the next time-period, 2004 and 2005, as the selected years and discards all others. This range and the following ranges were determined

during the interview process; in a later study, the effect of manipulating the year ranges will be examined.

27. If the medical researcher has no abstracts within the 2004-2005 periods, the procedure took abstracts from 2002 and 2003, as the selected years.
28. The classification of the medical researcher is based on a branching algorithm that examines the abstract classifications.
29. If all abstracts within the current period are the same, the medical researcher was labeled with the “field of study” associated with the abstracts. As an example if the medical researcher has three current abstracts, with each abstract labeled as “basic,” the researcher’s “field of study” is “basic.”
30. If the medical researcher has multiple abstract “field of study” values, then the classification procedure applies the domain expert’s rules for determining the medical researcher’s current “field of study.”
31. The test population of abstracts should be analyzed using the classification program results against the domain expert’s results to determine if the classification procedure is working correctly.
32. The “level of activity” is based on the total count of abstracts in the entire range. This dissertation defined “very active” as 12 or more abstracts over a 6.5 year period (2002- July 2008). Additionally, the medical researcher was considered “very active” if the majority of abstracts occurred in the current years, 2006 to July 2008.
33. The final analysis, using 101 randomly selected researchers and their abstracts, used the same procedure as the pilot study. The abstracts for the selected researchers were gathered and, starting from step 10 above, processed. No

changes were made in the classification source code or in the words and/or phrases selected for the pilot study.

7.10 OUTCOME MEASURES

7.10.1 Three outcome measures or endpoints

There are three outcome measures or endpoints in this dissertation:

- (a) The “abstract field of study” (AFOS) outcome measure is created by extracting the abstract from PubMed® and CRISP and categorizing the abstract into one of five categories: basic, outcomes, trial, translational, or nonstudy. The fifth additional classification, “nonstudy,” is included for those abstracts that cannot be identified as one of the four fields above.

- (b) The “researcher field of study” (RFOS) outcome measure is derived from analyzing each member's individual abstracts, evaluating the overall number for the current time-period and passing the categorized abstracts into a summation algorithm. The outcome measurement AFOS can have one of five categories: basic, outcomes, trial, translational, or nonstudy. The fifth classification, “nonstudy,” is included for those researchers that cannot be identified as one of the four fields above.

- (c) The “researcher level of activity and field of study” (RLAFS) outcome measure, is derived from analyzing each member's individual abstracts, evaluating the overall number for the current time-period, passing the categorized abstracts into a summation algorithm, and summing the abstract counts over the 6.5-year study period. The outcome measurement RLAFS can have ten possible categories: active basic, very active basic, active clinical outcome, very active clinical outcome, active clinical trial, very active trial, active translational, very active translational , active nonstudy or very active nonstudy.

The following operational definitions were provided by the domain expert and are considered the important characteristics for each “field of study.” The definitions differ depending on the location of the abstract (i.e., the abstract may be located in PubMed® or CRISP):

7.10.1.2 Basic Researcher

In PubMed®, this category refers to the investigator who uses a specialized laboratory to study in vitro samples (human or animal) or works on animal models.

In CRISP, this category suggests that the overall grant uses a specialized laboratory to study in vitro/in vivo samples (human or animal) or works on animal models

7.10.1.3 Clinical Outcomes Researcher

In PubMed®, this category refers to a medical researcher who follows patients and reports on their clinical outcomes.

In CRISP, this category suggests that the overall grant works with patients and uses routine clinical services in their investigations. The investigator's focus is on observational work that normally does not involve interventions (e.g., a randomized clinical trial). Some medical outcome abstracts may contain, on rare occasions, the randomization terms.

7.10.1.4 Clinical Trials Researcher

In PubMed®, this category refers to a medical researcher who works with patients and uses routine clinical services in their investigations. The medical researcher's focus is on experimental work that involves interventions (e.g., randomized clinical trial).

In CRISP, this category suggests that the overall grant works with patients and uses routine clinical services in their investigations. The investigator's focus is on experimental work that involves interventions (e.g., randomized clinical trial).

7.10.1.5 Clinical Translational Researcher

In PubMed®, this category refers to a medical researcher who is involved in experiments that include specialized laboratory procedures as well as clinical services.

In CRISP, this category suggests that the overall grant is involved in experiments that include specialized laboratory procedures as well as clinical services.

The translational researcher classification is derived from both the basic classification and any positive clinical classification (i.e., a positive clinical classification is either a clinical trial or clinical outcome finding).

7.10.2 Level of Activity

Once the abstract has been coded, the abstract dates were examined and a determination made as to the activity level of the medical researcher. The domain expert provides the operation definitions for “Active” and “Very Active” (see sections 7.8.2.1, 7.8.2.2). In this dissertation, the level of activity as presented by the domain expert is accepted as given. During the interview with the domain expert, a decision was made that a medical researcher who published two or more publications a year would be considered very active. The level of activity value has no correlation with the productivity of a researcher. These values were used simply as a starting point; a later study will examine this value and determine the effect on the classification procedure.

This study excludes all professionals within the Department of Medicine who have not published or have published as a co-author located anywhere but first or last within the abstract author field. The level of activity variable as used in this study does not examine the productivity of any single researcher, but is used as a first pass in the examination of group behavior. A very productive researcher who had never published

as first or last author would not appear in either level of activity category. A future study will expand the level of activity to include these co-authors.

7.10.2.1 Active Researcher

An activity number defined by the domain expert as having, over the last 6.5 years, publications or grants that total under 12.

7.10.2.2 Very Active Researcher

An activity number defined by the domain expert as having, over the last 6.5 years, publications or grants that totaled 12 or more. Additionally, the medical researcher was considered “Very Active” if the majority of abstracts occurred in the initial current years (2006-July 2008),

7.10.3 Current Time-Period

The domain expert has defined the investigators PubMed® or CRISP grant as “current” if the investigator has published over a recent 2.5-year block of time (2006- July:2008). In this dissertation, the ranges of values that make up the current variable as presented by the domain expert are accepted. The rationale behind this decision is that in medical research it may take 1-2 years for a paper or a grant to be accepted.

During the interview with the domain expert, a decision was made that a current time-period for a researcher would be a 2 to 2.5 year block of time. A decision was

made to set the most recent current time-period to 2.5 years, other time periods would consist of 2-year blocks of time. This dissertation is a first pass at addressing researcher classification; therefore, this dissertation accepted the period intervals as presented by the domain expert. A latter study will examine the current time-period values and determine their effect on the classification procedure.

Those who have not published or received grants in the current time range would be evaluated on their earlier 2-year block (2005-2006), or if no activity occurs in this period, on the earliest block (2002-2004). Any block of time not used is discarded. The block that is used is considered the current time-period.

7.11 OPERATIONAL DEFINITION FOR SUMMATION RULES

This section will introduce the rules used to determine researcher classification. The following rules should be considered a first pass at researcher classification, rather than specific rules to be followed exactly as written. These rules were discussed with the domain expert during the interview process and serve as a “best guess” for the classification procedure. This study will use the rules below as a starting point and later studies will determine the more exact formulas for each classification.

The 80/20 rule for basic researchers described below is strictly a rule-of-thumb, which recognized that basic researchers, in this domain, must work with physicians as part of their institutional duties. The majority of the Basic researcher’s work should be in

basic research, but a limited amount (i.e., defined here as 20 percent or less) of research may be in a clinical area.

From the discussion with the domain expert, the Clinical Trial and Outcome researcher publications were determined to differ from the Basic researcher, in that the clinical researcher may work directly with subjects who may also be patients. The interview suggested that the type of research performed by the Clinical Trial and Clinical Outcome researcher could overlap due to their clinical duties. In this study, there is an expectation that the abstract content between a clinical researcher (either trial or outcome) would not differ as broadly as the abstract content between the clinical researchers and a basic researcher. As a starting point, the decision was to use a simple majority (i.e., greater than 50%) as the cut-point for determining the number of abstracts that would determine the Clinical Trial or Clinical Outcome category. If the abstract counts are equal, then the classification procedure favored the Clinical Trial category. Later studies will examine this cut-point in more detail.

For the translational researcher, the domain expert and the information officer recognized that translational research is a relatively new categorization for researchers. Using the Basic researcher as a starting point, the opinion formed during the interview process was that an increase of more than 20% clinical work for a Basic researcher (either clinical trials or clinical outcomes) was sufficient to categorize that researcher as Translational. Because the translational research area is rooted in the laboratory, basic research is favored over clinical research. Later studies will examine this assumption in more detail.

This section of the dissertation emphasizes that the summation rules below are to be considered a starting point for a more general set of rules, which will be derived in a later study. The determination of the value of these rules is presented in the results section, which discusses the level of agreement between the classification procedure use of these rules, and the domain expert classification.

For the domain expert, an action folder was created, all abstracts were entered into the folder, and if the folder contains homogeneously labeled abstracts from within the current study period, the medical researcher was labeled with that “field of study.” For non-homogeneous abstracts contained within the folder, the following summation rules, with the caveats described above, derived from the initial discussion with the domain expert, were used to determine the medical researcher “field of study.”

The summation rules are not considered final; they are an attempt to develop a simple approximation of the domain expert’s method. Further study will be needed to determine if these rules can be enhanced. In addition, a single domain expert was used to develop the summation rules. Each domain expert provides their own rules for classification, it is unknown if the results would vary with another domain expert.

1. Summation rule for Basic Researcher

If 80% or greater of the abstracts “field of study” are Basic, the medical researcher is a Basic Researcher.

2. Summation rule for a Clinical Outcomes Researcher

If the majority (i.e., greater than 50%) of the research member's abstracts "field of study" are clinical outcomes then the medical researcher is a Clinical Outcomes Researcher.

3. Summation rules for a Clinical Trials Researcher

If the majority (i.e., greater than 50%) of the research member's abstracts are clinical trial abstracts then the medical researcher is a Clinical Trial Researcher.

4. Summation rules for a Clinical Translational Researcher

A decision was made to use the 80/20 rule for Translational research. If the research member has both Basic and clinical ((CLINICAL/ (CLINICAL + BASIC)) > 20) current abstracts and the clinical abstracts make up more than 20% of the current years' work, the medical researcher is a Translational Researcher. A latter study will examine the effect of this rule on the classification process.

5. If none of the above rules applies

Classify the medical researcher based on the majority abstract "field of study" for that medical researcher. If an abstract was not classifiable in a "field of study" for any reason (e.g., the abstract was missing) it was considered a

nonstudy abstract. If a researcher only had nonstudy abstracts, that researcher was classified as a nonstudy researcher.

7.11.1.1 Abstract field of study Chart

Each abstract had the following Analysis Chart [Figure 7-15] filled out by the domain expert. If blank, the abstract is labeled nonstudy.

Abstract Field of Study	
	<i>Selection</i>
BASIC	
Clinical Trial	
Clinical Outcomes	
Clinical Translational	

Figure 7-15 Abstract analysis chart

7.11.1.2 Level of activity and field of study chart

Each medical researcher had the following field of study chart [Figure 7-16] filled out by the domain expert. If blank, the researcher is labeled nonstudy.

Investigator Field of Study	<i>Level of Activity</i>	
	<i>Very Active</i>	<i>Active</i>
BASIC		
Clinical Trial		
Clinical Outcomes		
Clinical Translational		

Figure 7-16 Medical researcher field of study chart

The classification procedure uses a similar method, but creates arrays (i.e. an area of memory in a computer) that were used to hold information.

The outcome measures, the dependant variable, are what the classification procedure is trying to predict, in this case the “field of study” of the abstract-title variable and the “level of activity” and “field of study” of the medical researcher. The independent variables are publication time, study time, current years, triplet matching, and classification summarization.

7.12 SAMPLE SIZE CALCULATIONS

Tversky and Kahneman (1974) and other have suggested that sample size should always be carefully considered prior to running a study. The sample size for this dissertation is based on Table1 by Flack (1988, p 324), which suggests at least 100 medical researcher samples; this is at an expected power of 80% and an alpha equal to 0.5. The final study sample consisted of a random sample of 101 medical researchers randomly derived from the Department of Medicine phone book listing.

7.13 STATISTICS

Cohen (1960) introduced a coefficient of agreement in 1960, which was designed for nominal scales and measured rater versus chance agreement in the clinical-social-personality areas of psychology. Cohen's Kappa coefficient is a statistical measure of inter-rater reliability that takes into account the effect of chance agreement. It is considered a more robust measure than simple percent agreement calculations since Kappa factors in the agreement occurring by chance.

Kappa is defined (Cordes, 1994) as the ratio of the difference between obtained percent agreement and expected chance agreement to the difference between perfect agreement (i.e., 1.00) and expected chance agreement. Thus, Kappa indicates the extent to which obtained inter-rater agreement exceeds chance agreement. The purpose of Kappa is to determine how the raters agree; Kappa is not concerned with the relationship between the results and a "gold standard."

The equation for κ is:

$$K = (\text{Probability}(X) - \text{Probability}(Y)) / (1 - \text{Probability}(Y))$$

where $\text{Probability}(X)$ is the relative observed agreement among raters and $\text{Probability}(Y)$ is the probability that agreement is due to chance. If there is complete agreement among the raters then $\kappa = 1$. If there is no agreement among the raters (other than what is expected by chance) then $\kappa \leq 0$.

The Kappa value varies between 0, which indicates no agreement and 1, which indicates perfect agreement. The following strength of agreement table is taken directly

from Altman (1991, p 404) and provides information about the sub-ranges of the Kappa statistic (Figure 7-1). In the results section, the “strength of agreement” language used by Altman is also used to describe the relationship of Kappa.

Table 7-1 Strength of agreement

K Statistic	Strength of Agreement
Less than 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
Greater than 0.80	Very Good

SPSS version 16.1 was used to analyze the data for this dissertation a value of .40 or greater is considered significant.

7.14 STATISTICAL ASSUMPTIONS

The following three statistical assumptions are taken directly from Cohen (1960, p. 38):

1. The units are independent.
2. The categories of the nominal scale are independent, mutually exclusive, and exhaustive.
3. The judges operate independently.

For this dissertation, we also assume that there is no criterion for the "correctness" of judgments, and both human and computer are a priori deemed equally competent to make judgments. Additionally, there is no restriction placed on the distribution of judgments over categories for either categorization method (human or computer). Unlike Weighted Kappa, this Kappa also assumes a lack of order of the categories and the discrepancies between paired judgments are treated as equal to each other.

8.0 ANALYSIS OF RESEARCH QUESTIONS

The final study to answer the research questions used the same criteria as the pilot study. An abstract was linked to a researcher only when that researcher was the first or last author in the author field of the PubMed® or CRISP abstract. In the University of Pittsburgh, School of Medicine, Department of Medicine environment, it is generally accepted that the first author is the primary person conducting the works and the last author is the senior member of the working group. If the medical researcher was not the first or last author, the abstract was not used for analysis of that medical researcher. If the abstract is not classifiable (i.e., if the abstract does not fit into a category as defined by the domain expert), the abstract is placed into a nonstudy category. If the entire block of medical researcher's abstracts is not classifiable, the medical researcher was placed into a nonstudy category.

The primary outcome measure is the “researcher level of activity and field of study” (RLAFS) category, which is derived from analyzing each member's individual abstracts, evaluating the overall number for the current time-period, passing the categorized abstracts into a summation algorithm, and summing the abstract counts over the 6.5-year study period. The outcome measurement RLAFS can have ten possible categories: active basic, very active basic, active clinical outcome, very active

clinical outcome, active clinical trial, very active trial, active translational, very active translational , active nonstudy or very active nonstudy. Research question 3 compares the classification procedure and domain expert's categorization of the researcher into one of these ten categories.

The secondary outcome measure is “abstract field of study” (AFOS), which is created by extracting the abstract from PubMed® and CRISP and categorizing the abstract into one of five categories: basic, outcomes, trial, translational, or nonstudy. Research question 1 compares the classification procedure and domain expert's categorization of the abstract into one of these five categories.

The third outcome measure is the “researcher field of study” (RFOS) category, which is derived from analyzing each member's individual abstracts, evaluating the overall number for the current time-period and passing the categorized abstracts into a summation algorithm. The outcome measurement RFOS can have five possible categories: basic, outcome, trial, translational, or nonstudy. Research question 2 compares the classification procedure and domain expert's categorization of the researcher into one of these five categories.

8.1 PILOT STUDY

The pilot study, which consisted of 16 randomly selected de-identified medical researchers and their associated 87 abstracts, was used to determine the ability of the

program to categorize the researchers and their abstracts. To determine the effect of randomness on the population, the pilot study compared computer abstract classification to a random abstract classification. Kappa showed no agreement with the random paring either with the pilot abstract population ($K = 0.006$, $n = 87$) or against the pilot researcher population ($Kappa = 0.007$, $n = 16$).

The modifiers “moderate” and “good,” which are used to describe Kappa, are taken from Altman’s “Strength of Agreement” table (Figure 7-18) and are associated with a range of Kappa values, 0.41-0.60 and 0.61 – 0.80, respectively. In the pilot study, Kappa showed good agreement between the automatic and domain expert classification for the abstract’s field of study ($Kappa = 0.685$, $n = 87$). Kappa also showed good agreement between the automatic and domain expert classification of the medical researcher’s “level of activity and field of study” ($Kappa = 0.628$, $n = 16$). “Level of activity” and “field of study” were not evaluated during the pilot study because of the low numbers of abstracts and researchers used.

8.2 FULL STUDY

The full study, which consisted of 101 randomly selected de-identified medical researchers and their associated 504 abstracts, was used to determine the ability of the program to categorize the researchers and their abstracts. The domain expert read (over a period of weeks) each researcher’s abstract or abstracts, classified each abstract, and then made a determination of the researcher’s classification based on the

abstract's classification. The classification procedure methods were identical to those used in the pilot study.

8.3 INITIAL POPULATION

The initial population consisted of 564 professionals (e.g., MD, PhD) within the University of Pittsburgh, School of Medicine, Department of Medicine. Of these, 285 had published at least one article in PubMed® (Figure 8-1) and 145 had a notice of grant award in CRISP (Figure 8-2). Of the 564 researchers, 305 had either published a PubMed® abstract and/or had a CRISP abstract (Figure 8-3).

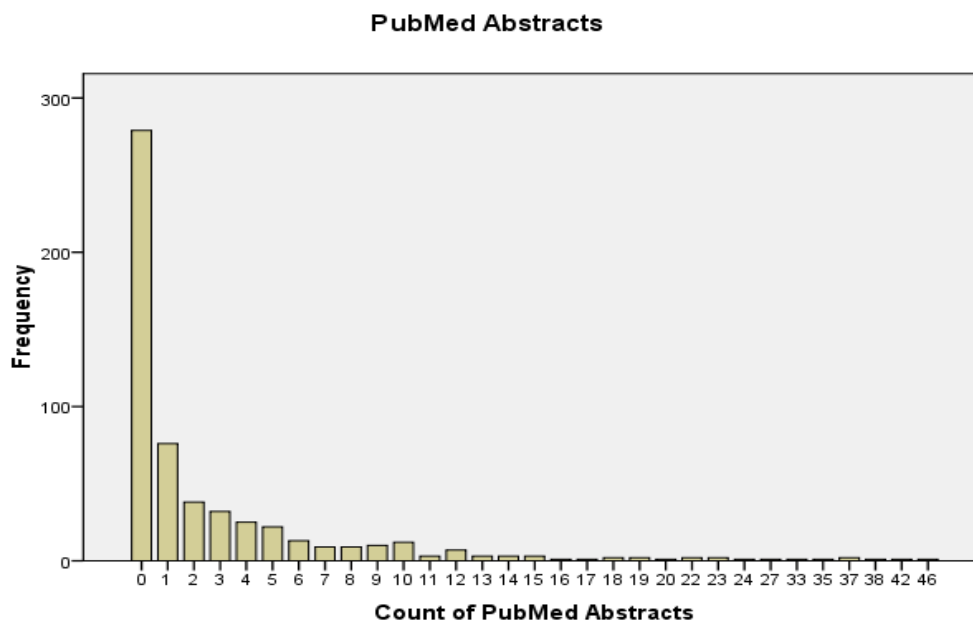


Figure 8-1 Medical Researchers who have published at least one PubMed® Abstract

Figure 8-1 shows that a little over half of the professionals (285/564, 50.5%) in the Department of Medicine were found to have at least one abstract in the time period studied.

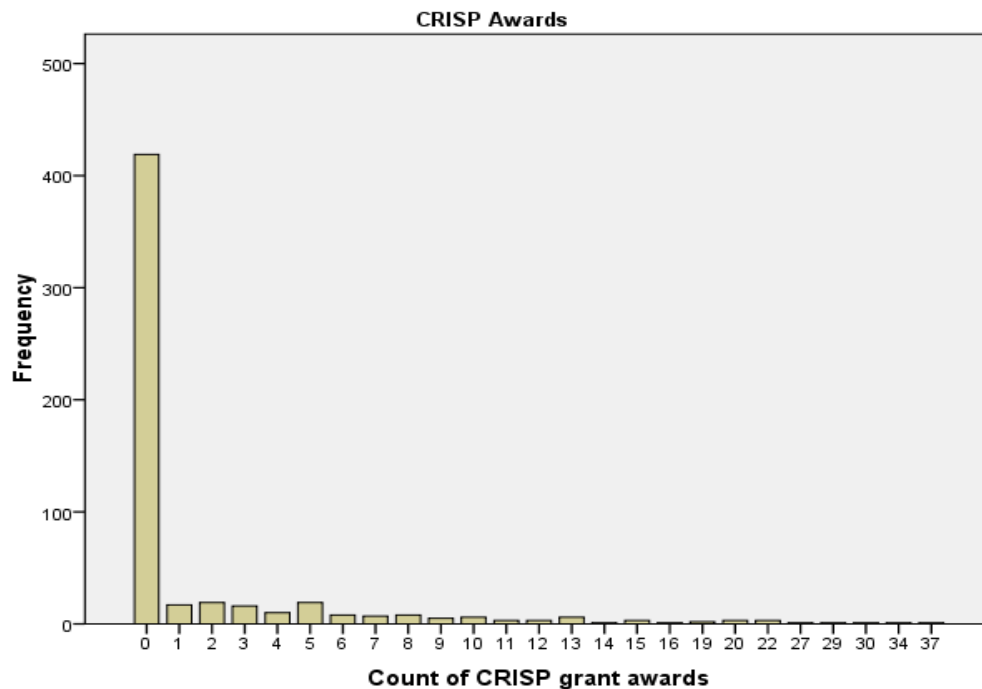


Figure 8-2 Medical Researchers who have at least one CRISP abstract

Figure 8-2 shows that a little over twenty-five percent of the professionals (145/564, 25.7%) in the Department of Medicine where found to have at least one CRISP abstract in the time period studied.

An analysis of the distribution of “level of activity” is presented in Figure 8-3 as a power-law curve for the overall distribution of investigators within the Department of Medicine domain. The finding of power-law curves as a function of author seniority is not unexpected and has been discussed by others (De Solla Price, 1963; MacRoberts &

MacRoberts, 1982). The blue portion of the graph represents abstracts extracted from PubMed® and the stacked red portion of the graph represents abstracts extracted from CRISP over the full time period 2002-July 2008.

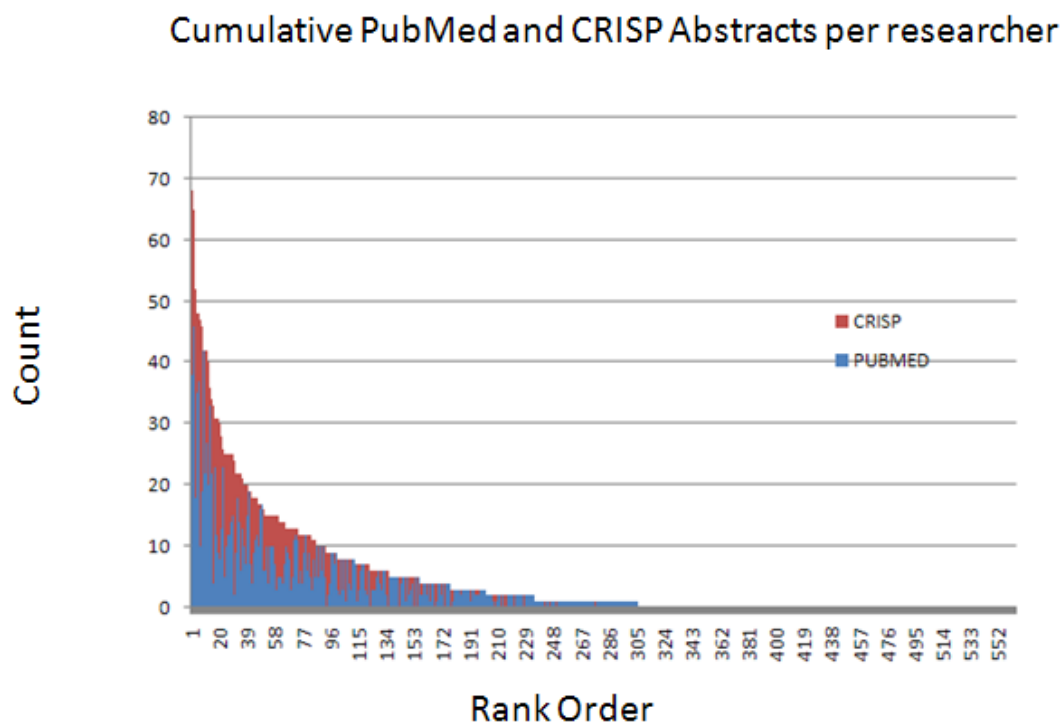


Figure 8-3 Researcher published PubMed® abstracts and/or CRISP abstract

The abscissa is in rank order and a cumulative count of 12 or more indicated a very active researcher and a count of less than 12 indicated an active researcher. The graph shows that 54.1% (305/564) of the professionals within the DOM domain publish as first or last author and/or are primary investigators in a CRISP grant.

8.4 ACTIVE AND VERY ACTIVE CLASSIFICATION

A determination of active versus very active based on their cumulative PubMed® and CRISP abstracts per researcher count was used to ascertain their activity level. Those with less than 12 (the cutoff value suggested by the domain expert) were considered active; those equal to or having more than 12 were classified as very active. The breakdown is shown in the following table (Table 8-1).

Table 8-1 Count of active versus very active researcher

		ACTIVITY LEVEL	
		Frequency	Percent
Valid	ACTIVE	58	57.4
	VERY ACTIVE	43	42.6
	Total	101	100.0

8.5 RESULTS OF THE RESEARCH QUESTIONS

8.5.1 Can an automated procedure classify abstracts from academic medical researchers' publications and grants into a "field of study?" What is the level of agreement between the automated procedure and the results derived from a domain expert?

In research question 1, 504 documents were processed and analyzed for chance pairing (i.e., each expert or procedure value was paired with a random table generated value). A Kappa value of -.020 was found for the classification procedure processed abstracts (Table 2) and a value of .001 was found for the expert processed abstracts (Table 3) versus random classification. Both values are sufficiently close to zero to suggest that classification by the procedure and by the expert were not random.

Table 8-2 Classification procedure versus random pairing – Question 1

		Symmetric Measures			
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	-.020	.021	-.937	.349
N of Valid Cases		504			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 8-3 Kappa for domain expert versus random pairing – Question 1

		Symmetric Measures			
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.001	.022	.038	.970
N of Valid Cases		504			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

An analysis of the inter-rater agreement (Table 4) of the 504 documents by the classification procedure and by the domain expert (Kappa = 0.535, Table 5) suggests moderate agreement level between both methods. (EXPERT refers to Domain Expert and CLASSIFICATION refers to the automatic classification procedure)

Table 8-4 Classification procedure versus domain expert – Question 1

		EXPERT * CLASSIFICATION					
Count							
		CLASSIFICATION					
		BASIC	NONSTUDY	OUTCOME	TRANSLATIONAL	TRIAL	Total
EXPERT	BASIC	137	8	6	11	2	164
	NONSTUDY	11	46	24	2	0	83
	OUTCOME	1	2	107	12	7	129
	TRANSLATIONAL	21	6	30	25	10	92
	TRIAL	0	0	19	4	13	36
	Total	170	62	186	54	32	504

Table 8-5 Kappa value - Question 1

		Symmetric Measures			
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.535	.027	22.132	.000
N of Valid Cases		504			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The highest agreements between the classification procedure and domain expert occurred in the Basic abstracts (83.5% agreement [137/164]) and outcomes (82.9% agreement [107/129]). The lowest agreement occurred with translational abstracts (27.2% agreement [25/92]), while agreements with trial and nonstudy abstracts were moderate (36.1% agreement [13/36] and 55.4% agreement [46/83] respectively).

8.5.2 Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their “field of study?”
What is the level of agreement between the automated procedure and the results derived from a domain expert?

In research question 2, 101 researchers' classifications, for five categories, were processed and analyzed for chance pairing in a manner described in section 8.5.1. A Kappa value of -.01 was found for the classification procedure processed researcher classification (Table 6) and a value of .026 was found for the expert processed researcher classification (Table 7) versus random classification. Both values are

sufficiently close to zero to suggest that researcher categorization by the classification procedure and by the domain expert were not random.

Table 8-6 Classification procedure versus random pairing - Question 2

		Symmetric Measures			
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	-.001	.049	-.023	.982
N of Valid Cases		101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 8-7 Domain expert versus random pairing - Question 2

		Symmetric Measures			
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	-.026	.045	-.545	.586
N of Valid Cases		101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

An analysis of the inter-rater agreement of the 101 researchers by the classification procedure and by the domain expert method of classification (Kappa = .572) suggests a moderate agreement level between both methods.

Table 8-8 Classification procedure versus domain expert - Question 2

EXPERT * CLASSIFICATION							
Count							
		CLASSIFICATION					
		BASIC	NONSTUDY	OUTCOME	TRANSLATIONAL	TRIAL	Total
EXPERT	BASIC	22	0	0	5	1	28
	NONSTUDY	1	2	1	0	0	4
	OUTCOME	2	0	29	6	0	37
	TRANSLATIONAL	6	0	2	15	0	23
	TRIAL	1	0	3	3	2	9
	Total	32	2	35	29	3	101

Table 8-9 Kappa value - Question 2

Symmetric Measures					
		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.572	.062	9.508	.000
N of Valid Cases		101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The highest agreements between the classification procedure and domain expert occurred in the basic abstracts (78.6% agreement [22/28]) and outcomes (78.4% agreement [29/37]). The lowest agreement occurred with trial abstracts (22.2%

agreement [2/9]), while agreements with translational and nonstudy abstracts were moderate (65.2% agreement [15/23] and 50.0% agreement [2/4] respectively).

The methods used to determine a basic and outcome researcher appears to have a high level of success. The improvement in translational classification is probably due to the summation process used by the procedure used to determine a translational researcher.

8.5.3 Can an automated procedure use multiple categorized abstracts from an individual medical researcher to classify that individual into their “level of activity and field of study?” What is the level of agreement between the automated procedure and the results derived from a domain expert?

Question 3 is designed to answer the more complex questions referenced in the early sections of this dissertation. Questions such as:

1. In section 4.1.1, “*Solving the National Institutes of Health mentoring problem*,” research question 3 describes how the classification procedure can be used to identify mentors. Question 3 serves to identify those who are "very active" in their field and could potentially act as mentors of those who are only "active" or just starting in their field.

2. In Section 4.1.2, "*Laying the foundation for library based push technology*," question 3 now describes how the classification procedure could be used to filter information to medical researchers and retain the most active members. In reference to filtering, those who have published frequently in their field (i.e., those researcher who are "very active" in their field) should have already examined the information in their field than those who are just entering that field, (i.e., those researchers would be "active" in their field) because of their larger experience in that field. Therefore, the "very active" researchers would have a "user profile" that would suggest very specific information as compared to those who are only "active" in their field.

In research question 3, 101 researchers' classifications, for ten categories, were processed and analyzed for chance pairing in a manner described in section 8.5.1. A Kappa value of -.018 was found for the classification procedure processed abstracts (Table 10) and a value of .007 was found for the domain expert processed abstracts (Table 11). Both values are sufficiently close to zero to suggest that researcher categorization into ten categories by the classification procedure and by the domain expert were not random.

Table 8-10 Classification procedure versus random pairing - Question 3

Symmetric Measures		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	-.018	.033	-.513	.608
N of Valid Cases		101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 8-11 Domain expert versus random pairing - Question 3

Symmetric Measures		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.007	.035	.209	.835
N of Valid Cases		101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

An analysis of the ten category inter-rater agreement of the 101 researchers by the classification procedure and by the domain expert method of classification (Kappa = 0.634) suggests good agreement level between both methods.

Table 8-12 Classification procedure versus domain expert - Question 3

EXPERT * CLASSIFICATION

		CLASSIFICATION									
		ACTIVE BASIC	ACTIVE NONSTUDY	ACTIVE OUTCOME	ACTIVE TRANSLATIONAL	ACTIVE TRIAL	VERY ACTIVE BASIC	VERY ACTIVE OUTCOME	VERY ACTIVE TRANSLATIONAL	VERY ACTIVE TRIAL	Total
EXPERT	ACTIVE BASIC	14	0	0	1	1	0	0	0	0	16
	ACTIVE NONSTUDY	1	2	1	0	0	0	0	0	0	4
	ACTIVE OUTCOME	0	0	22	4	0	0	0	0	0	26
	ACTIVE TRANSLATIONAL	4	0	1	1	0	0	0	0	0	6
	ACTIVE TRIAL	0	0	2	3	1	0	0	0	0	6
	VERY ACTIVE BASIC	0	0	0	0	0	8	0	4	0	12
	VERY ACTIVE OUTCOME	0	0	0	0	0	2	7	2	0	11
	VERY ACTIVE TRANSLATIONAL	0	0	0	0	0	2	1	14	0	17
	VERY ACTIVE TRIAL	0	0	0	0	0	1	1	0	1	3
	Total	19	2	26	9	2	13	9	20	1	101

Table 8-13 Kappa value - Question 3

Symmetric Measures		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.634	.053	15.105	.000
	N of Valid Cases	101			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The highest agreements between the classification procedure and domain expert with the researchers occurred in the active basic (87.5% agreement [14/16]), active outcome (84.6% agreement [22/26]), and very active translational (82.4% agreement [14/17]). The lowest agreement occurred with active translational abstracts (16.7% agreement [1/6]), active trial (16.7% agreement), and very active trial (33.3% agreement [1/3]). Agreement with active nonstudy (50.0% agreement [2/4]), very active basic (66.7% agreement [8/12]), and very active outcome (63.6 % agreement [7/11]) were moderate. Very active nonstudy was not found in either the classification procedure method or domain expert method.

8.6 PILOT STUDY VERSUS FINAL STUDY

The pilot study suggested good inter-rater agreement between the classification procedure and domain expert. A slight decrease was noted in the kappa values for the pilot study versus the final study for the abstract field of study classification (0.685, 0.535 respectively) and a slight increase for researcher level of activity and field of study

classification (0.628, 0.634 respectively). This may reflect the effect of randomly selecting a population that is slightly different from the pilot study population.

8.7 STUDY CONCLUSION

The objective of this dissertation was to propose that an informationist, using simple library concepts (i.e., Resource Description Frame), established library functions (i.e., simple PubMed® queries), and informatics tools (i.e., text mining) can develop automated tools to assist a research institution rationalize infrastructure resource requirements needed by their faculty based on simple categorization, which is derived from their research focus. This dissertation suggested a tool, the classification procedure, which used a simple dialog between the informationist and a domain expert as a means to classify researchers into their level of activity and field of study. This analysis was designed to test whether the classification procedure performs as well as a domain expert in understanding the academic makeup of the academic medical institution.

The ability of the classification program to identify translational abstracts in research question 1 was low, only 27.2 % of the translational abstracts were identified by the classification procedure. In research question 2, the classification procedure was able to identify the translational researcher, with 65.2% of the translational researchers identified by the classification procedure. In research question 3, when level of activity

was introduced, agreement of the very active translational researchers rose to 82.4%. Interestingly, in research question 3, the identification of the active translational researcher dropped with a comparison of only a 16.7% agreement as compared to the domain expert. The results from research question 3 suggest that the classification procedure is unable to identify the beginning translational researcher with high accuracy.

The worst researcher classifications occurred in research question 2, where trial classifications only had a 22.2% agreement with the domain expert. In research question 3, the lowest comparisons occurred with the active trial (16.7% agreement) and very active trial (33.3% agreement) as compared to the domain expert.

The analysis does show a high level of inter-rater correlation in the basic and outcome categories in all three-research questions. This may be due to the makeup of the School of Medicine, Department of Medicine researchers with their heavy emphasis on teaching and NIH grants, which stress basic research, and the large internal medicine environment, which stress outcomes research. The nonstudy category (ignored for the researcher classification analysis) does suggest that the non-classified abstract or researcher is similarly identified in both the classification procedure and domain expert methods.

The purpose of this study was to determine how well the categorization procedure performs as compared to a single domain expert in classifying medical researchers into their field of study and level of activity. The study did show that the Kappa values from the classification procedure versus domain expert were higher than

the Kappa values derived from comparing the classification procedure versus random classification.

Other studies will explore additional features of the researcher's publication (e.g., co-authors, key words or mesh terms, full text analysis) and try to improve upon Kappa. This study does suggest that a rapid method of classifying a medical research domain is possible by using the services of the information officer and domain expert. The advantages and disadvantages of the classification procedure versus the manual method are presented in table 14.

Table 8-14 Advantages and Disadvantages of the Classification Procedure

Advantages	Disadvantages
Fast	Not as accurate at the manual domain expert method
Easy to develop	Dependent on a single domain expert rather than a panel of domain experts.
Information sources are public databases	Information officer must have experience as a programmer
Inexpensive to develop, operate, and maintain.	There may be concerns with privacy and security issues in maintaining the researcher database.

It would be interesting in a later study to compare this method to other methods (e.g., subject terms assigned in indexes or citation analysis) in classifying medical researchers. Since each method uses a different part of the publication, synergy between all three methods may also lead to an improvement in classification.

This study and the Kappa values derived from it, serve as an exploratory step in the classification process. A later study will have to determine what methods would improve the classification of the researchers, especially the active translational, very active trial, and active trial researchers.

9.0 FUTURE DIRECTION AND CONCLUSION

9.1 CLASSIFICATION AS A METHOD OF DETERMINING A MEDICAL RESEARCHER'S LIBRARY NEEDS

An article by Talja (2002), who presents insight into information sharing within an academic community, can be used to suggest a future direction for the classification procedure. The author suggests that senior researchers use an informal, socially-based method of seeking information, while junior researchers tend to use a more formal structured method of seeking information. The classification procedure by breaking the population within the academic community into those that are more active (i.e., those who tend to be more senior) and those who are less active (i.e., those who tend to be more junior) can help provide a determination of the need for formal information seeking resources (e.g., databases, space for library users, information content) within the community.

There is an interesting paragraph in the article by Talja (p. 8), where the author discusses the role of librarians in information sharing in an academic community. The author suggests that researchers prefer to collaborate with those professionals that speak the same language as they do. The author further noted that many researchers

were reluctant to use the general technical skills that typically are presented by the library professional. The researchers, who did use a particular librarian for searches, explained that these librarians possessed specific training in their specialties, understood the terms that used, and are considered “qualified” for reference searching.

Classification, as presented in this dissertation, could serve as an exploratory step in understanding the language of the medical researcher, especially the senior medical researcher, providing a dictionary of commonly used terms specific to that research field.

Table 9-1 Domain Landscape

LEVEL	FIELD OF STUDY			
		CE	TOTAL	
ACTIVE	BASIC	14	16	87.50%
VERY ACTIVE	BASIC	8	12	66.67%
ACTIVE	NONSTUDY	2	4	50.00%
VERY ACTIVE	NONSTUDY	0	0	0
ACTIVE	OUTCOME	22	26	84.62%
VERY ACTIVE	OUTCOME	7	11	63.64%
ACTIVE	TRANSLATIONAL	1	6	16.67%
VERY ACTIVE	TRANSLATIONAL	14	17	82.35%
ACTIVE	TRIAL	1	6	16.67%
VERY ACTIVE	TRIAL	1	3	33.33%
		70	101	69.31%
Kappa = .634				

A very interesting and useful application of the classification procedure would be to describe the research subgroup landscape of the domain. An example of a “domain landscape”, derived from this study, is presented in Table 9.1. The categories are displayed along with their relative percents within the domain. This broad outline could be used to formulate a strategy with the purpose of improving services to the domain members. A domain landscape could be used as an unbiased starting point for discussions of resource allocation. The domain landscape could be stratified by years, which would provide a method of looking at historical subgroup population trends and predicting future direction. A domain landscape, used as a service from the library, could be used as a method of recruiting a junior medical researcher by demonstrating that the subgroup of interest has an active community with many published members to draw upon. Or, the domain landscape could be used as a method of recruiting a senior medical researcher by demonstrating the subgroup of interest has “room to grow.”

The counts in Table 9.1 suggest that the landscape pattern in this domain consists of a larger group of outcome researchers, followed by a slightly smaller group of basic researchers, a smaller group of translational researchers, and finally a very small group of clinical trial researchers. This table is derived from a random sample of the domain and the group percentages should be similar when the classification system is applied to the entire domain. This pattern may suggest that services that interest outcome, basic, and translational researchers would be of high interest to this domain. The smaller clinical trial group might find an advantage in working with the library services of another domain, such as the University of Pittsburgh Cancer Institute, which may have a larger number of clinical trial researches.

The nonstudy category may be used as an indicator of the level of understanding that the domain expert has in reference to the research makeup of the domain. In Table 9.1, the nonstudy category is very small with only four researchers, suggesting that the four categories given by the domain expert provide an adequate fit to this domain. If the nonstudy category is large, then this may suggest that the domain expert understanding of this part of the institution's research population may not be sufficient.

Another use of the domain landscape is to analyze specific subgroups within the institutional domain. As an example, in Table 9.1, the translational researchers' levels of activity patterns are different from the outcome and basic levels of activity patterns. The information from Table 9.1 suggests that those who publish very actively (using this study's current value of 12 or more abstracts over six years) are a larger group than those who publish less than 12 abstracts over a six year period. The reverse is true for the basic and outcomes researcher. This information may be used to propose more expensive library services to the very active translational researchers, who should have the resources to fund these services (i.e., the CTSA grant was designed to assist the well-published translational researcher). As discussed earlier, the value of 12 or more abstracts is not used to suggest the productivity of any researcher. It is simply a first step in dividing the researcher population into two groups. A later study will look at the level of activity and determine how to take into account co-authorships and publications in other non-PubMed® databases.

The domain landscape could be used in collection development or to find the pivot group within the domain. As described in this dissertation, the translational researchers are a blend of basic and clinical research. The library services could be

directed to the translational and basic researchers or the translational and clinical (outcomes, trial or both) researchers. The translational researchers, used as the pivot group, could act as the champions for library services in either basic or clinical groups.

The library could offer the above a service that would provide a data dictionary of the “language of the researcher” within that domain or present the “domain landscape” to a new department head. The department head would get a quick overview of the divisions within the department that could be used in the discussion with the various division heads that make up that department. The domain landscape could be used to describe people resources (e.g., the large number of very active translational researchers would be able to mentor the grant recipients) that could be contributed to a grant.

9.2 PREDICTING ACCEPTANCE OF A NEW SERVICE

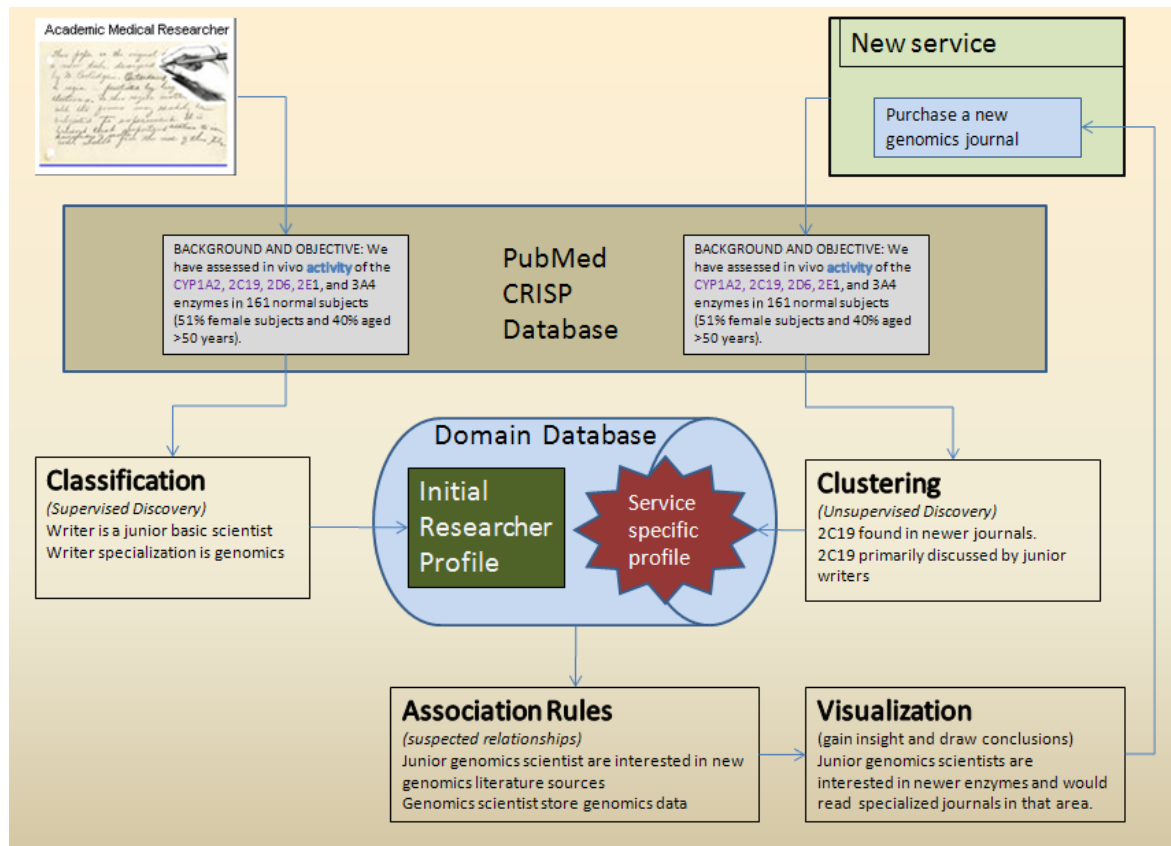


Figure 9-1 Future direction: Predictive instrument

Another future direction would use the classification process as part of a larger predictive instrument, to determine acceptance or removal of services (or as discussed later, any information object) within an academic medical domain, without the necessity of asking each medical researcher his or her opinion. This would save time, expense,

and provide the academic medical administration a method of dynamically evaluating many services for addition or removal without causing faculty response fatigue.

The predictive instrument, using a structure described by Chen and Liu (2004), is outlined in Figure 9-1 and consists of the following:

1. PUBMED® & CRISP DATABASE: The process begins when the medical researcher identifies their research interest by storing their abstracts and grants, respectively, in PubMed® and CRISP. Other sources may be mined in the future (e.g. CiteSeerX) and included as source information. These abstracts are retrieved from their public sites using web services and stored within their own domain specific database.
2. CLASSIFICATION: The classification system, as described in this PhD dissertation, is a supervised text mining process that uses both PubMed® and CRISP abstracts to organize the medical researcher's interests into specific categories. In this dissertation, the researchers were classified into level of activity and field of study. Other classification systems, created by the informationist and domain expert, can also be developed with each classification system serving a specific purpose within the domain.
3. DOMAIN DATABASE: The classification system produces an automatic initial researcher profile for the medical investigator, which is stored within a domain-specific database. This initial profile is a starting point, a way of rapidly gathering user-specific information about the medical researcher without the need for direct

discussions. This profile can also be static (i.e., the current research interest) or dynamic (i.e., a series of classification snapshots that define the medical researcher over past, current, or future time) depending on the needs of the domain. This researcher profile can also be combined with surveys or other methods used to evaluate members of a domain. The domain database will also contain information specific to the new service.

4. **NEW SERVICE:** In the example above, the domain administration has asked the informationist to determine if a new service, in this case the purchase of a new genomics journal, would be considered important by any member of the domain. Outside of library services, the informationist could evaluate whether a medical research domain would have an interest in the development of a new genomics laboratory or a new information-gathering tool.
5. **CLUSTERING:** A clustering process, also known as exploratory data analysis (EDA), is an unsupervised process that determines trends, correlations, and patterns within the data. Words and/or phrases associated with the new journal (e.g., enzymes, 2C19 - which is an enzyme linked to bladder cancer) are extracted from the PubMed® and CRISP database and, in this example; a determination of frequency of “2C19” use found in the new journal is compared to the frequency of “2C19” use found in the medical researcher’s abstracts.

6. ASSOCIATION RULES: During this process, expected relationships are explored and discussed, i.e., in the figure 9.1, the informationist has determined that genomic researchers of a junior level appear more interested in using this tool to acquire new information in “2C19” research than senior basic researchers who gain the same information via their social networks (e.g., scientific meetings, reviewing papers).
7. VISUALIZATION: The informationist interactively draws conclusions or gains insight into the data using various graphical techniques such as two or three dimensional graphs or multi-dimensional charts (e.g., Chernoff plots). In our example, the informationist has arrived at the conclusion that the junior genomics scientist frequency counts of the “2C19” enzyme are correlated. The informationist suspects that this population would consider the new journals important.
8. The process repeats for other potential new services within the domain or for discarding old, unused services. This predictive instrument can be used within a single domain (i.e., Department of Medicine), within a virtual medical research team (i.e., Specialized Center of Clinically Oriented Research (SCCOR) in Pediatric and Adult Pulmonology), or even against an entire university.

9.3 DYNAMIC CLASSIFICATION

The classification procedure described in this dissertation is a static picture of the medical researcher. The classification procedure can also be used dynamically to create a moving picture of that researcher's career. The academic medical institution could use this dynamic method of predicting those who might be interested in a particular developing grant or research area without requesting numerous intrusive surveys or interviews.

9.4 FURTHER INVESTIGATION

Within the Department of Medicine, specialties exist (e.g., Cardiology, Pulmonary, Hematology/Oncology) and an interesting area to examine is the effect of the subspecialties on the classification procedure. The question that could be asked is whether applying classification within the domain subspecialties and using the same procedure framework within the subspecialties would provide greater inter-rater agreement.

In fact, one could also examine the professional social networks that exist within the subspecialties, replacing the formal departmental subspecialties, and determine if classification inter-rater agreement is affected. Mathiak and Eckstein (2004, p.48) make the observation that "each area of research develops its own vocabulary," which may suggest that social networks, developed around specific projects, develop their own

language and may be better classification predictors. These vocabularies may have a temporal component (i.e. the researchers would work together on a particular project for a few years, publish, and then go on to another project with another social network) that could be identified by the classification procedure.

Another area of investigation is to determine if the classification procedure will work outside of its initial testing domain. Within the University of Pittsburgh, the School of Medicine has many other departments (e.g., Anesthesiology, Cell Biology & Physiology, Psychiatry, Physical Medicine & Rehabilitation, Radiation Oncology, Radiology), which have similar public abstract storage domains. In fact, the classification procedure could be used to analyze other research institutions or even other academic fields, determining their research structure, which then provides insight into their specific information needs. The library could offer classification procedure as a billable service to an institutional technology group, providing an estimate of the possible users of their product. An example of another academic field is Systems Biology, which may benefit from this method of classification.

9.5 FINAL CONCLUSION

Dr. Herbert S. White, in his article (1988) "*Oh, Why (and Oh, What) Do We Classify?*" makes two interesting observation in reference to classification of material within the library. One, "there is no such thing as *the* library, there is only *my* library and what it

contains (p.43, para. 6).” and two, classification fulfills “our responsibilities as information professionals to help users get what they need (p. 43, para. 7). “

This dissertation suggests that classification of research professional may be one way of expanding the library by creating a specific view of the library based on the career choice of the library user, the academic medical researcher, moving towards White’s concept of “my library.” Understanding what category the research user occupies may provide insight to the information professional, giving the information professional the ability to “help users get what they want.”

This dissertation attempted to demonstrate that fusion of library science and informatics techniques could provide services that would be very difficult to perform using traditional human methods. The classification procedure, since it uses material found within the public sector as its source of information, is inexpensive to operate and maintain.

The particular profession (i.e., academic medical research) described within this dissertation is a niche population, specific to a highly professional career. An informationist, who has knowledge of these domains, is in a position to use inexpensive public resources to assist the institutional and library administration in understanding the fundamental structure of this researcher population. If the fundamental structure of the domain is understood, novel new services specific to that domain can be introduced, funded and maintained.

This dissertation suggests that the informationist, who has training in both library and informatics fields, is in the best position to work with the medical research institution to develop these new services.

10.0 ADDENDUM A. IRB APPROVAL

Cecchetti, Alfred

From: irb+@pitt.edu [irb+@pitt.edu] **Sent:** Thu 10/2/2008 9:10 AM
To: Cecchetti, Alfred
Cc:
Subject: PI Notification: The IRB determined your project met the criteria for an exemption
Attachments:



University of Pittsburgh *Institutional Review Board*

3500 Fifth Avenue
Pittsburgh, PA 15213
(412) 383-1480
(412) 383-1508 (fax)
<http://www.irb.pitt.edu>

Memorandum

To: ALFRED CECCHETTI, PhD
From: CHRISTOPHER RYAN, PhD, Vice Chair
Date: 10/2/2008
IRB#: [PRO08090037](#)
Subject: An automatic Library Information Science based method of classifying medical researchers into domain specific subgroups

The above-referenced project has been reviewed by the Institutional Review Board. Based on the information provided to the IRB, this project includes no involvement of human subjects, according to the federal regulations [§46.102(f)]. That is, the investigator conducting research will not obtain data through intervention or interaction with the individual, nor will obtain identifiable private information. Should that situation change, the investigator must notify the IRB immediately. Given this determination, you may begin your project.

If any modifications are made to this project, please contact the IRB Office to ensure it continues to meet the no human subjects determination.

Upon completion of your project, be sure to finalize the project by submitting a termination request. Please be advised that your research study may be audited periodically by the University of Pittsburgh Research Conduct and Compliance Office.

11.0 ADDENDUM B. CLASSIFICATION CODE

11.1 FUNCTION FLOW LIST

```
// Load the investigators
LoadInvestigator(); // CLEAN AND Load PubMed® ID and
SEARCHNAME from THEINVESTIGATORS where PUBMED > 0

LoadCRISPInvestigator(); // Load CRISP ID and SEARCHNAME from
THEINVESTIGATORS where CRISP > 0

LoadGRANTINFO(); // Preload the Grant information from the
GRANTINFO table

LoadCRISP(); // Load the CRISP information from CRISP0208
where ID > 0
```

```
StartEnd("START");
```

```
// Garbage collection
GC.Collect();
GC.WaitForPendingFinalizers();
```

```
this.textBox1.Font = new Font("Arial", 14);
```

```
// tell me what I am doing
textBox1.AppendText("Extracting sentences from PUBMED" +
    Environment.NewLine);
```

```
// Set the ACTIVE and VERY ACTIVE COUNTERS
```

```

CreateActivityLevel();

// Load PUBMED information for all records
LoadPUBMED("ALL");

// load of the words that I will look at
LoadSEARCHWORDS("FIELD OF STUDY");

this.textBox1.Font = new Font("Arial", 6);

        // load search phrases
        // LoadSEARCHPHRASES();

// Make all the sentences
Make_PUBMEDSentences();

// insert the sentences into table THESENTENCES
InsertSentences();

// UPDATE the table PUBMED with the FIELD OF STUDY;
updatePUBMEDCount();

// Load PUBMED information, WHICH NOW HAS COUNT
INFORMATION
LoadPUBMED("ALL"); // Get Pubmed data from PUBMED
TABLE - put into memory objects

// Update THE INVESTIGATOR TABLE
updateTHEINVESTIGATORCount();

this.textBox1.Font = new Font("Arial", 12);

// Reload the investigator array
ReLoadInvestigator();

// now create my FIELD OF STUDY variables for PUBMED.
CreateFIELD OF STUDY PUBMED();

StartEnd("END");

```

11.2 SELECTED FUNCTIONS

```
private void PUBMEDObjectGetFieldOfStudy(string anewvalue, int aPubMedIndex,
string aWorkingOn)
{
    // This associates the FIELD OF STUDY back to the document

    // Just to be certain that we do not have any spaces or blanks
    // The results should only be BASIC CLINICAL or TRANSLATIONAL
    // anewvalue is what
    // Insert directly into array
    // anewvalue = what we just found
    // aDocNumber = what document we are working on
    // aWorkingOn = what field MESH, TITLE or ABSTRACT

    anewvalue = anewvalue.Trim();

    string inObjectNow = "";

    if (aWorkingOn == "TITLE")
    {
        inObjectNow = PUBMEDObject[aPubMedIndex].TITLEFIELD OF STUDY;

        PUBMEDObject[aPubMedIndex].TITLEFIELD OF STUDY =
replaceObject(inObjectNow, anewvalue);

    } else if (aWorkingOn == "MESH")
    {
        inObjectNow = PUBMEDObject[aPubMedIndex].MESHFIELD OF STUDY;
        PUBMEDObject[aPubMedIndex].MESHFIELD OF STUDY =
replaceObject(inObjectNow, anewvalue);

    } else if (aWorkingOn == "ABSTRACT")
    {
        inObjectNow =
PUBMEDObject[aPubMedIndex].ABSTRACTFIELD OF STUDY;
        PUBMEDObject[aPubMedIndex].ABSTRACTFIELD OF STUDY =
replaceObject(inObjectNow, anewvalue);

    } else
    {
```

```

        MessageBox.Show (" Wow.. we never should have gotten here" +
            Environment.NewLine +
            "aWorking on = " + aWorkingOn + Environment.NewLine +
            "New value " + anewvalue );
    }

}

private void CreatePUBMEDCRISPFIELDOFSTUDYPUBMED()
{
    string myquery = "";
    double basic_clinical = 0;
    double xBASIC_PLUS_CLINICAL = 0;
    double xCLINICALPERCENT = 0.000;
    int xCLINICALCOUNT = 0;
    string xresearch = "";

    // ABSTRACT
    for (int x = 0; x < HoldObject.Count; x++)
    {
        Application.DoEvents();

        basic_clinical = HoldObject[x].BASIC_PI +
            HoldObject[x].CLINICALTRANSLATIONAL_PI +
            HoldObject[x].CLINICALOUTCOMES_PI +
            HoldObject[x].CLINICALTRIAL_PI;

        xCLINICALCOUNT = HoldObject[x].CLINICALOUTCOMES_PI +
            HoldObject[x].CLINICALTRIAL_PI +
            HoldObject[x].CLINICALTRANSLATIONAL_PI;

        xresearch = "B:" + HoldObject[x].BASIC_PI.ToString() +
            "O:" +
            HoldObject[x].CLINICALOUTCOMES_PI.ToString() +
            "TR:" + HoldObject[x].CLINICALTRIAL_PI.ToString()
            +
            "T:" + HoldObject[x].CLINICALTRANSLATIONAL_PI;

        if (basic_clinical == 0)
        {
            HoldObject[x].AFIELDOFSTUDY = "NOTCLASSIFIED";
        }
    }
}

```

```

else if ((HoldObject[x].BASIC_PI > 0) && (xCLINICALCOUNT >
0))
{
    // MAKE IT HARDER TO BECOME A TRANSLATIONAL RESEARCHER
    // USE THE 80 / 20 RULE
    // YOUR CLINICAL WORK MUST BE GREATER THAN 20 % OF YOUR
TOTAL WORK
    // CLINICAL/(CLINICAL + BASIC) > 20

    xBASIC_PLUS_CLINICAL = HoldObject[x].BASIC_PI +
xCLINICALCOUNT;
    xCLINICALPERCENT = (xCLINICALCOUNT /
xBASIC_PLUS_CLINICAL);

    // this may be a better indicator of translational

    xCLINICALPERCENT = ((double)xCLINICALCOUNT / (double)
HoldObject[x].BASIC_PI);

    /*
    MessageBox.Show("basic " + HoldObject[x].BASIC_PI +
Environment.NewLine +
        "Clinical " + xCLINICALCOUNT + Environment.NewLine +
        "basic + clinical " +
xBASIC_PLUS_CLINICAL.ToString() +
        Environment.NewLine +
        "xclinicalpercent " + xCLINICALPERCENT.ToString());

    */

    // if (xCLINICALPERCENT > xTRANSLATIONALCUTOFF )
    // if nothing is in this column then make it basic, if
something is
    // in this field then make it translational.
    if (HoldObject[x].CLINICALTRANSLATIONAL_PI > 0)
    {
        HoldObject[x].AFIELDOFSTUDY = "TRANSLATIONAL";
        // +xresearch + " xC% " +
xCLINICALPERCENT.ToString();
    }
    else
    {
        HoldObject[x].AFIELDOFSTUDY = "BASIC";
        // +xresearch + " xC% " +
xCLINICALPERCENT.ToString(); ;
    }

}
else if ((HoldObject[x].BASIC_PI > 0) &&
(HoldObject[x].CLINICALTRIAL_PI == 0) && (HoldObject[x].CLINICALOUTCOMES_PI
== 0) )
{
    HoldObject[x].AFIELDOFSTUDY = "BASIC";

```

```

        }
        else if (HoldObject[x].CLINICALOUTCOMES_PI >
HoldObject[x].CLINICALTRIAL_PI)
        {
            HoldObject[x].AFIELDOFSTUDY = "CLINICAL OUTCOMES";
        }
        else if (HoldObject[x].CLINICALTRIAL_PI >=
HoldObject[x].CLINICALOUTCOMES_PI)
        {
            HoldObject[x].AFIELDOFSTUDY = "CLINICAL TRIAL";
        }
        else
        {
            HoldObject[x].AFIELDOFSTUDY = "UNKNOWN";
        }

        myquery = "UPDATE PUBMEDCRISPTHEINVESTIGATORS SET " +
            "FIELDOFSTUDY_PI = '" +
                HoldObject[x].AFIELDOFSTUDY + "'," +
            "FINALFIELDOFSTUDY_PI = ACTIVITYLEVEL + '" + " " +
                HoldObject[x].AFIELDOFSTUDY + "'" +
            "WHERE ID = " + HoldObject[x].ID.ToString();

        TheSQLQuery(myquery);

        textBox1.AppendText(myquery + Environment.NewLine);

    }

    /*
    *
    *

        else if (HoldObject[x].CLINICALTRANSLATIONAL_PI > 0)
        {
            HoldObject[x].AFIELDOFSTUDY = "TRANSLATIONAL";
        }
        else if ((HoldObject[x].BASIC_PI > 0) &&
(HoldObject[x].CLINICALTRIAL_PI > 0) )
        {
            HoldObject[x].AFIELDOFSTUDY = "TRANSLATIONAL";
        }
        else if ((HoldObject[x].BASIC_PI > 0) &&
(HoldObject[x].CLINICALOUTCOMES_PI > 0))
        {
            HoldObject[x].AFIELDOFSTUDY = "TRANSLATIONAL";
        }

    *
    * */
    }

```

12.0 BIBLIOGRAPHY

1. Adusumilli, P., Chan, M., Ben-Porat, L., Mullerad, M., Stiles, B., Tuorto, S., & Fong, Y. (2005). Citation characteristics of basic science research publications in general surgical journals. *Journal of Surgical Research*, 128(2), 168-173.
2. Aerts, D., Broekaert, J., & Mathijs, E. (1999). Einstein meets Magritte: An interdisciplinary reflection: The White Book of "Einstein Meets Magritte."
3. Altman, D. G. *Statistics for Medical Research*. London: Chapman & Hall; 1991.
4. Andrews, J.E., Pearce, K.A., Ireson, C., & Love, M. M., (2005, April). Information-seeking behaviors of practitioners in a primary care practice-based research network (PBRN). *Journal Medical Library Association*, 93(2), 206–212.
5. Angell, M. (1986). Publish or perish: A proposal. *Annals of Internal Medicine*, 104(2), 261-262.
6. Beaver, D. (2001). Reflections on Scientific Collaboration (and its Study): Past, Present, and Future - Feature Report. *Scientometrics*, 52(3), 365-377.
7. Berner, E.S. (2008, May-June). Implementation Challenges for Clinical and Research Information Systems: Recommendations from the 2007 Winter Symposium of the American College of Medical Informatics. *Journal of the American Medical Informatics Association*, 15(3), 281-282.
8. Bertolami, C.N. (2008). President-Elect's address. *Journal of Dental Education*, 72(7), 758-759.
9. Bhattacharjee, Y. (2007). Postdoctoral training NSF, NIH emphasize the importance of mentoring [News of the Week]. *Science*, 317, 1016b. Retrieved on 10 October 2008 from <http://www.sciencemag.org/cgi/content/full/317/5841/1016b>.
10. Black box. (2008). In Merriam-Webster Online Dictionary. Retrieved September 29, 2008, from [http://www.merriam-webster.com/dictionary/black box](http://www.merriam-webster.com/dictionary/black%20box).
11. Brennan, M. J., Hurd, J. M., Blečić, B. D., & Weller, A. C. (2002, November). A snapshot of early adopters of e-journals: Challenges to the library. *College & Research Libraries*, 63(6), 515-526.

12. Buehring, G. C., Buehring, G. E., & Gerard, P. D. (2007). Lost in citation: Vanishing visibility of senior authors. *Scientometrics*, 72(3), 459–468.
13. Calvert, J. (2006). What's special about basic research? *Science, Technology, & Human Values*, 31(2), 199-220.
14. Calvert, J., & Martin, B. R. (2001). Changing conceptions of basic research? *Background Document for the Workshop on Policy relevance and Measurement of Basic Research*, 29-30.
15. Campbell, M., & Plumb, G. (2003). Generic names for soft drinks by county. Retrieved on 17 September 2008 from <http://www.popvssoda.com:2998/countystats/total-county.html>.
16. Carpenter, S. (2007). Carving a career in translational research. *Science*, 317, 966-967.
17. Chen, S. Y., & Liu, X., (2004). The contribution of data mining to information science, *Journal of Information Science*, 30(6), 550-558.
18. Chu, W.W., Johnson, D.B., & Kangarloo, H. (1999). A medical digital library to support scenario and user-tailored information retrieval. *IEEE Transactions on Information Technology in Biomedicine*, 4(2), 97-107.
19. Classify. (2008). In Merriam-Webster Online Dictionary. Retrieved September 23, 2008, from <http://www.merriam-webster.com/dictionary/classify>.
20. Clemmons, N. W., & Clemmons, S. L. (2005). Five years later: medical reference in the 21st century. *Medical Reference Service Quarterly*, 24(1), 1-18.
21. Cockburn, I., & Henderson, R, (1996, November 12). Public-private interaction in pharmaceutical research. *Proceeding of the National Academy of Sciences U S A*, 93(23), 12725–12730.
22. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
23. Cohen, K. B., & Hunter, L. (2004). Natural language processing and systems biology. In W. Dubitzky & F. Azuaje (Editors). *Artificial Intelligence Methods and Tools for Systems Biology* (pp.147–173) Springer, Norwell, MA. Retrieved on 1 March 2009 from <http://www.springerlink.com/content/l3414833xq665j89/>.
24. Cohen, K. B., & Hunter, L., (2008). Getting started in text mining. *PLoS Computational Biology*, 4(1), 1-3.
25. Cole, S. S. (2006). Researcher behavior that leads to success in obtaining grant funding: A model for success. *Research Management Review*, 15 (2),1-16. Retrieved on 2 February 2009 from http://www.ncura.edu/content/news/rmr/docs/research_behavior.pdf.

26. Cordes, A. (1994, April). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech & Hearing Research*, 37(2), 264-279.
27. Cox, B. J., Wessel, I., Norton, G. R., & Swinson, R. P. (1994). Citation patterns in anxiety disorders research in 14 journals: 1990–1991. *American Journal of Psychiatry*, 151(6), 933-936.
28. Crowley, G. H., Leffel, R., Ramirez, D., Hart, J. L., & Armstrong, T. S. (2002). User perceptions of the library's web pages: A focus group study at Texas A&M University. *Journal of Academic Librarianship*, 28, 205–10.
29. Crowley, W. F. (2003, April). Translation of basic research into useful treatments: How often does it occur? *The American Journal of Medicine*, 114(6), 503-505.
30. Davidson, J. R., & Middleton, C. A. (2006). Networking, networking, networking: The role of professional association memberships in mentoring and retention of science librarians. *Science & Technology Libraries*, 27(1), 203-224.
31. Decker, R. & Höppner, M. (2006). Strategic planning and customer intelligence in academic libraries, *Library High Tech News Information*, 24 (4), 504-14.
32. Detlefsen, E. G. (2002). The education of informationists, from the perspective of a library and information sciences educator. *Journal of the Medical Library Association*, 90 (1), 59-67.
33. Detlefsen, E. G. (2004). The clinical research informationist. *Reference Services Review*, 32 (1), 26–30.
34. Drenth, J. P. H. (1998). Multiple authorship: The contribution of senior authors. *Journal of the American Medical Association*, 280, 219-221.
35. Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16, 681–735.
36. Edmunds, A., & Morris, A. (2000). Problem of information overload in business organizations: A review of the literature. *International Journal of Information Management*, 20(1), 17-28.
37. English, K. (2003, July/August). The changing landscape of leadership. *Research Technology Management*, 46(4), 9-11.
38. Ewigman, B. (2008). CTSAs and family medicine research--time to get connected. *The Annals of Family Medicine*, 6(2), 181-182.
39. Eysenbach, G. (2000, October). The impact of preprint servers and electronic publishing on biomedical research. *Current Opinion in Immunology*, 12(5), 499-503.

40. Faloutsos, C., & Oard, D. (1995). A survey of information retrieval and filtering methods, technical report. University of Maryland Computer Science Dept, College Park, MD.
41. Figg, W. D., Chau, C.H., Okita, R., Preusch, P., Tracy, T.S., McLeod, H., Reed, M., Pieper, J., Knoell, D., Miller, K., Speedie, M., Blouin, R., Kroboth, P., Koda-Kimble, M., Taylor, P., & Cohen, J. (2008). Pharm. D. Pathways to Biomedical Research: The National Institutes of Health Special Conference on Pharmacy Research. *Pharmacotherapy*, 28 (7), 821.
42. Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two-rater kappa statistic. *Psychometrika*, 53(3), 321-325.
43. Franklin, M., & Zdonik, S. (1998, June). Data in your face: push technology in perspective. Paper presented at in proceedings of the ACM SIGMOD Conference on Management of Data, Seattle, Washington. 516-519.
44. Garfield, E. (1979a), Perspective on citation analysis of scientists, Chapter 10. In book *Citation Indexing. Its Theory and Application in Science, Technology, and Humanities*. Originally published in 1979 by John Wiley & Sons, Inc., New York, NY. Reprinted in 1983 by ISI Press, Philadelphia, PA. pp. 240-252. Retrieved 30 September 2008 from <http://www.garfield.library.upenn.edu/ci/chapter10.pdf>.
45. Garfield, E. (1979b). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1, 359-375.
46. Gralla, R. J. (1999). Recommendations for the use of antiemetics: Evidence-based, clinical practice guidelines. *Journal of Clinical Oncology*, 17 (9), 2971-2994.
47. Green, D. W. (1986). Writing, jargon, and research. *Written Communication*, 3, 364-381.
48. Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, 43(5-6), 907-928. Retrieved on 17 September 2008 from <http://tomgruber.org/writing/onto-design.pdf>.
49. Gustitus, C. (1998). The push is on: What push technology means to the special librarian. *Information Outlook*, 21-24.
50. Guyatt G., Cook D., & Haynes B. (2004). Evidence based medicine has come a long way. *British Medical Journal*, 329, 990-991.
51. Hallmark, J., & Lembo, M. F. (2003, spring). Leaving science for LIS: interviews and a survey of librarians with scientific and technical degrees. *Issues in Science and Technology Librarianship* [serial online], 37. Retrieved 2 October from <http://istl.org/03-spring/refereed1.html>.

52. Hart, J. L., Vicki Coleman, V., & Yu, H. (2000). Marketing electronic resources and services: Surveying faculty use as a first step. *The Reference Librarian*, 67/68, 41-55.
53. Hersh, W. (2002). Medical informatics education: An alternative pathway for training informationists. *Journal of the Medical Library Association*, 90(1), 76–9.
54. Herubel, J. P. V. M., & Buchanan, A. L. (1994). Citation studies in the humanities and social sciences: a selective and annotated bibliography. *Collection Management*, 18, 89–137.
55. Hill, S., & Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. *ACM SIGKDD Explorations Newsletter*, 5(2), 179-184.
56. Hoskisson, T. (1997). Making the right assumptions: know your user and improve the reference interview. *The Reference Librarian*, 59, 67-75.
57. Houlihan, R. (2005). The academic library as congenial space: More on the Saint Mary's experience, *New Library World*, 106, 1208/1209, 7-15.
58. Hripcsak, G., & Heitjan D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35, 99-110.
59. Hurd, J. M., Blecic, B. D., & Vishwanatham, R. (Jan. 1999). Information Use by molecular biologists: Implications for library collections and services. *College & Research Libraries*, 60, 31-43.
60. Hust, A. (2005). Query expansion methods for collaborative information retrieval. *Informatik-Forschung und Entwicklung*, 19(4), 224–238.
61. Ichise, R., Takeda, H., & Ueyama, K. (2006). Exploration of researcher' social network for discovering communities. Japanese Society for Artificial Intelligence (JSAI) 2005 Workshops, LNAI 4012, 458-469.
62. Ioannidis, J. P. (2004). Materializing research promises: opportunities, priorities and conflicts in translational medicine. *Journal of Translational Medicine*, 2, 1-5. Retrieved on July 25, 2008 from <http://www.translational-medicine.com/content/2/1/5>.
63. JAMA Instructions for authors. Journal of the American Medical Association. Retrieved 1 July 2008 from <http://jama.ama-assn.org/misc/ifora.dtl#Abstracts>.
64. Joswick, K. E., & Stierman, J. K. (1997). The core list mirage: A comparison of the journals frequently consulted by faculty and students. *College & Research Libraries*, 58, 48–55.
65. Kademani, B. S., & Kalyane, V. L. (1998). Scientometric portrait of R. Chidambaram, the Indian Nuclear Physicist, based on citation analysis. *Kerala Library Professionals' Organization (KELPRO) Bulletin*, 2(1), 13 –29.

66. Kavulya, J. M. (2004). Marketing of library services: a case study of selected university libraries in Kenya. *Library Management*, 25, 118–26.
67. Kendall, K. E., & Kendall, J. E. (1999). Information delivery systems: An exploration of web pull and push technologies. *Communications of the AIS*, 1(14) 1-41.
68. Klink, S. (2004). Improving document transformation techniques with collaborative learned term-based concepts. In A. Dengel, M. Junker, & A. Weisbecker (editors) *Reading and Learning*, volume 2956 of Lecture Notes in Computer Science, 281–305. Springer.
69. Knoben, J. E, Phillips, S. J., Snyder, J.W., & Szcur, M. R. (2004). The National Library of Medicine and Drug Information. Part 2: An evolving future. *Drug Information Journal*, 38, 171-180.
70. Korjonen-Close, H. (2005, June). The information needs and behaviour of clinical researchers: a user-needs analysis. *Health Information and Libraries Journal*, 22, 96-106.
71. Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43, 27-43.
72. Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, O. E., & Ramirez, A. M. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52, 1148–1156.
73. Kostoff, R. N. (2002). Text mining for global technology watch. *Encyclopedia of Library and Information Science*, Marcel Dekker, Inc., New York, NY, in press.
74. Kuhn, T. (1962). *The Structure of Scientific Revolutions*, Chicago University Press, Chicago.
75. Kwasik, H., Fulda, P. O., & Ische, J. P. (2006). Strengthening professionals: a chapter-level formative evaluation of the Medical Library Association mentoring initiative. *Journal of the Medical Library Association*, 94 (1), 19-29.
76. Laine-Cruzel, S., Lafouge, T., Lardy, J. P., & Abdallah, N. B. (1996). Improving information retrieval by combing user profiles and document segmentation. *Information Processing and Management*, 32(3), 305–315.
77. Lam, W., Mukhopadhyay, S., Mostafa, J., & Palakal, M. (1996). Detection of shifts in user interests for personalized information filtering, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 317-325, August 18-22, Zurich, Switzerland.
78. Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *Library Quarterly*, 66, 161-93.

79. Lepori B., Canton L., & Mazza, R. (2002). Push communication services: a short history, a concrete experience and some critical reflections. *Studies in Communication Sciences*, 2(1), 149-164.
80. Littman. H., Di Mario, L., Plebani, M. & Marincola, F. M. (2007). What's next in translational medicine? *Clinical Science*, 112, 217–227.
81. MacLean, M., Anderson, J., & Martin, B. R., (1998). Identifying research priorities in public sector funding agencies: mapping science outputs to user needs. *Technology Analysis & Strategic Management*, 10(2), 139–155.
82. MacRoberts, M. H., & MacRoberts, B. R. (1982, August). A re-evaluation of Lotka's Law of scientific productivity. *Social Studies of Science*, 12(3), 443-450.
83. MacRoberts, M. H., & MacRoberts, B.R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information*. 40(5), 342-349.
84. McConchie, A. (2002). The Great Pop vs. Soda Controversy. Retrieved on 17 September 2008 from <http://www.popvssoda.com:2998/>.
85. Magrill, R. M. & St. Clair, G. (1990). Undergraduate term paper citation patterns by disciplines and level of course. *Collection Management*, 12(3/4), 25–56.
86. Mao, J. (2002). Translational pain research: Bridging the gap between basic and clinical research. *Pain*, 97, 183–187.
87. Mathiak, B., & Eckstein, S. (2004). Five steps to text mining in biomedical literature. Paper presented at Data mining and text mining for bioinformatics European workshop, 2004. Retrieved on 20 March 2009 from http://www2.informatik.hu-berlin.de/Forschung_Lehre/wm/ws04/7.pdf.
88. March, R. W. Jr. (2008, February). Text mining patent literature: A case scenario of a methodology for analyzing unstructured data. *Searcher*, 16(2), 30-5.
89. Marcum, J.W. (2003). Visions: the academic library in 2012, *D-Lib Magazine*, 9(5). Retrieved on 1 October 2008 from <http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/may03/marcum/05marcum.html>.
90. Marriott, R. (2002). Yes, but how do we know if it's working? Evidence regarding impact on clinical practice of access for health service staff to bibliographic databases and full text electronic journals. *Library Review*. 51(7), 358 – 363.
91. Medical Library Association. Our mission. [Web document]. Chicago, IL: The Association. [24 Oct 2000; cited 30 September 2008]. Retrieved from <http://www.mlanet.org/about/mission.html>.
92. Miller, E. J. (2001). An introduction to the resource description framework. *Journal of Library Administration*, 34(3/4), 245 – 255.

93. Morrison, L. (2008). The CTSAs, the Congress, and the scientific method, *Journal of Investigative Medicine*, 56, 7-10.
94. Morville, P. (2005). *Ambient Findability: What We Find Changes Who We Become*. Sebastopol, CA: O'Reilly Media.
95. National Institutes of Health, Mentor roles and responsibilities. Retrieved 30 September 2008 from <http://internships.info.nih.gov/mentor.html>.
96. Neill, S. D. (1989). The information analyst as a quality filter in the scientific communication process. *Journal of Information Science*, 15, 3-12.
97. Nwogu, K.N. (1997). The medical research papers: structure and functions. *English for Specific Purposes*, 16, 119–138.
98. Ontario Neurotrauma Foundation. Glossary. Retrieved October 1, 2006 from <http://www.onf.org/knowledge/glossary.htm>.
99. Pardridge, W. M. (2003). Translational science: what is it and why is it so important? *Drug Discovery Today*, 8(18), 813-815.
100. Peritz, B. C. (1994). On the heuristic value of scientific publications and their design; a citation analysis of some clinical trials. *Scientometrics*, 30, 175–186.
101. Petrelli, D., Hansen, P., Beaulieu, M., Sanderson, M., Demetriou, G. & Herring, P. (2004). Observing users - Designing clarity: A case study on the user-centred design of a cross-language retrieval system. *Journal of the American Society for Information Science and Technology*, 55 (10), 923-934.
102. Phelan, T. J. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45, 117–136.
103. Pitkin, R. M. (1987). The importance of the abstract [Editorial]. *Obstetrics and Gynecology*, 70(2), 267-269.
104. Pober, J. S., Neuhauser, C. S., & Pober, J. M. (2001). Obstacles facing translational research in academic medical centers. *Federation of American Societies for Experimental Biology*, 15, 2303-2313.
105. Porter, A. L. (1977). Citation analysis: queries and caveats. *Social Studies of Science*. 7(2), 257-267.
106. Porter, S. R., & Umbach, P. D. (2001). Analyzing faculty workload data using multilevel modeling. *Research in Higher Education*, 42, 171–196.
107. Powell, R. R., Baker, L. M., & Mika, J. J. (2002). Library and information science practitioners and research, *Library & Information Science Research*, 24(1), 49-72.
108. Price, D. J. D. (1963). *Little Science, Big Science*. Columbia University Press, New York. 118 pp.

109. Price, D., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, 21, 1011-1018.
110. Quandt, S. A., & Arcury, T. A. (1997). Qualitative methods in arthritis research: Overview and data collection. *Arthritis Care & Research*, 10(4), 273-281.
111. Rankin, J. A., Grefsheim, S. F., & Canto, C. C. (2008). The emerging informationist specialty: a systematic review of the literature. *Journal of the Medical Library Association*, 96(3), 194-208.
112. Ried K., Farmer E. A., & Weston K. M. (2006) Setting directions for capacity building in primary health care: A survey of a research network BMC Family Practice, 7(8). Retrieved 16 September 2008 from <http://www.biomedcentral.com/1471-2296/7/8>.
113. Rosenberg, L. (1999). Physician–scientists — endangered and essential. *Science*, 283, 331-332.
114. Rosenberg, S. A. (1996). Secrecy in medical research. *The New England Journal of Medicine*, 334(6), 392-394.
115. Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L., & Brown, S. H. (2008). A Model for evaluating interface terminologies, *Journal of the American Medical Informatics Association*, 15(1), 65-76.
116. Rossi, L. (2006). \$83.5 million NIH grant to University of Pittsburgh establishes institute for clinical and translational research. Retrieved on October 4, 2006 from <http://newsbureau.upmc.com/TX/ReisGrant.htm>
117. Rothman, R.A. (1979). Occupational roles: power and negotiation in the division of labor. *The Sociological Quarterly*, 20(4), 495-515.
118. Sambunjak, D., Straus, S.E, & Marusic, A. (2006). Mentoring in academic medicine: a systematic review. *Journal of the American Medical Association*, 296, 1103–1115.
119. Sathe, N. A., Jerome R, & Giuse, N. B. (2007). Librarian-perceived barriers to the implementation of the informationist/information specialist in context role. *Journal of the Medical Library Association*, 95(3), 270–274.
120. Shapiro, D. W. (1994). The contributions of authors to multiauthored biomedical research papers. *The Journal of the American Medical Association*, 271(6), 438 - 442.
121. Schein, E. H. (1993). On dialogue, culture, and organizational learning. *Organizational Dynamics*, 22(2), 40– 51.
122. Shine, K. I. (1998). Encouraging clinical research by physician scientists. *Journal of the American Medical Association*, 280(16), 1442-1444.

123. Shosteck, H. & Fairweather, W. P. (1979, summer). Physician response rates to mail and personal interview surveys. *The Public Opinion Quarterly*, 43(2), 206-217.
124. Shultz, S. M. (1996). Medical jargon: ethnography of language in a hospital library. *Medical Reference Services Quarterly*, 15(3), 41-47.
125. Shurin, S. B. (2008). Clinical Translational Science Awards: Opportunities and challenges. *Clinical and Translational Science*, 1(1), 4-4.
126. Snyderman, R. (2004). The clinical researcher— An “emerging” species. *Journal of the American Medical Association*, 291(7) (Reprinted), 882-883.
127. Sonnenwald, D.H, & Pierce, L.G. (2000, May). Information behavior in dynamic group work contexts: interwoven situational awareness, dense social networks and contested collaboration in command and control. *Information Processing and Management: an International Journal*, 36(3), 461-479.
128. Sung, N. S., Crowley, W. F., & Genel, M. (2003). Central challenges facing the national clinical research enterprise. *Journal of the American Medical Association*, 289, 1278-1287.
129. Talja, S. (2002), Information sharing in academic communities: types and levels of collaboration in information seeking and use. *New Review of Information Behavior Research*, 3, 143-160.
130. Talja, S, & Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases. A domain analytic study in four scholarly disciplines. *Journal of Documentation*. 59(6), 673-691.
131. Tan, T. (2007, October 29). Mining legacy content with OCR. *Publishers Weekly*, 254(43), 20-20.
132. Turati, C., Usai, A., & Ravagnani, R. (1998). Antecedents of co-ordination in academic international project research. *Journal of Managerial Psychology*, 13(3-4), 188-198.
133. Tversky, A & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.
134. UCSF, School of Pharmacy (2003) Glossary- basic science research. University of California, San Francisco. Retrieved on October 7, 2006 from <http://pharmacy.ucsf.edu/glossary/r/>.
135. US Department of Health & Human Services, National Institutes of Health, Office of Extramural Research, Q&A and FAQs, question number 16. Retrieved 30 September 2008 from <http://grants1.nih.gov/training/q&a.htm#mentor>.
136. Ventura, O. N., & Mombrú, A. W. (2006). Use of bibliometric information to assist research policy making. A comparison of publication and citation profiles of full and

- associate professors at a school of chemistry in Uruguay. *Scientometrics*, 69(2), 287-313.
137. Von Schneidemesser, L. (1996). Soda or Pop? *Journal of English linguistics*, 24 (4), 270 - 287.
 138. Vucovich, L. A., Baker, J. B. & Smith, J. T. (2008). Analyzing the impact of an author's publications. *Journal of the Medical Library Association*, 96(1), 63–66.
 139. Weinberg, A. M. (1967). *Reflections on Big Science*. The M.I.T. Press, Cambridge, MA. 182 pp.
 140. Weise, F. J. (2004). Being there: the library as place. *Journal of Medical Library Association*, 92(1), 6-13.
 141. Whelan, N. (2007, Spring/Summer). Knowledge is power: Center for Clinical and Translational Informatics working to make “Bench to Bedside” a reality. *Informatics Today*. Retrieved October 28, 2007 from www.dbmi.pitt.edu.
 142. White, H. S. (1988, June 15). Oh, why (and oh, what) do we classify? *Library Journal*, 11, 42-43.
 143. White, H. D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure? Longitudinal evidence from the "GloboNet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology*, 55(2), 111-126.
 144. Will, N. (2006). Data mining: improvement of university library services, *Technological Forecasting and Social Change*, 73, 1045–1050.
 145. Woolf S. H. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association*, 299, 211-213.
 146. Wu, Y.D., & Liu, M. (2001). Content management and the future of academic libraries, *Electronic Library*, 19(6), 432-439.
 147. Zerhouni, E. A. (2005a). Translational and clinical science-time for a new vision. *New England Journal of Medicine*, 353(15), 1621-1623.
 148. Zerhouni, E.A. (2005b). US biomedical research: Basic, translational, and clinical sciences. *Journal of the American Medical Association*, 294, 1352–1358.
 149. Zerhouni, E. A. (2006). Translational and clinical science. *New England Journal of Medicine*. 354(9). 978-979, Response to Fox, R. J. (2006) Translational and Clinical Science. *New England Journal of Medicine*, 354(9), 978.
 150. Zerhouni, E. A. (2007). Translational research: moving discovery to practice. *Clinical Pharmacology & Therapeutics*, 81, 126–128.